Proceedings of the International Workshop on

Augmenting User Models
with Real World Experiences to Enhance
Personalization and Adaptation
(AUM)

co-located with the International Conference on User
Modeling, Adaptation and Personalization
July 15, 2011, Girona, Spain

Editors: Fabian Abel$^1$, Vania Dimitrova$^2$, Eelco
Herder$^2$, Geert-Jan Houben$^1$

$^1$ Web Information Systems, Delft University of Technology
Mekelweg 4, 2628 Delft, the Netherlands
{f.abel,g.j.p.m.houben}@tudelft.nl
$^2$ School of Computing, University of Leeds, UK
V.G.Dimitrova@leeds.ac.uk
$^3$ L3S Research Center, Leibniz University Hannover, Germany
herder@l3s.de
Table of Contents

Preface .................................................................................................................. i
    Fabian Abel, Vania Dimitrova, Eelco Herder and Geert-Jan Houben

Semantically Enriched Machine Learning Approach to Filter YouTube Comments for Socially Augmented User Models ...... iii
    Ahmad Ammari, Vania Dimitrova and Dimoklis Despotakis

The Personal Adaptive In-Car HMI: Integration of External Applications for Personalized Use ........................................ iv
    Sandro Rodriguez Garzon and Mark Poguntke

A Semantic Approach to Extract Individual Viewpoints from User Comments on an Activity ........................................... v
    Dimoklis Despotakis, Lydia Lau and Vania Dimitrova

Recommender Systems and the Social Web .............................. vi
    Amit Tiroshi, Tsvi Kuflik, Judy Kay and Bob Kummerfeld

Towards a Digital Learner Identity .................................................... vii
    Adriana Berlanga and Peter Sloep

Core Aspects of Affective Metacognitive User Models ........... viii
    Adam Moore, Victoria Macarthur and Owen Conlan

Extraction of Professional Interests from Social Web Profiles ... ix
    Fabian Abel, Eelco Herder and Daniel Krause
Preface

The digital world, i.e. our interaction with computer systems, becomes more and more connected with the physical world, i.e. our real-world activities and experiences. This changes the way we use technologies and opens up new opportunities for personalization and adaptation. People blog, post, chat, comment, tweet about things that matter to them: what they had for dinner, what their job activities were, what they thought about a particular television broadcast, et cetera. People share content about their activities, e.g. pictures taken at a concert, videos of business meetings, reports on business trips, personal stories. This abundant digital information stream has become an important backchannel in our daily lives. We constantly create digital traces about our experiences, which can be invaluable source for personalization.

The time is ripe for developing new adaptation paradigms that exploit digital traces to extend users’ personalized experience by connecting the digital, social and physical worlds. Hence, traditional adaptation mechanisms (such as feedback, help, guidance) can be extended to become more effective by taking into account not only the user’s experience in the digital world (i.e. the conventional user modeling paradigm), but also relevant experience (of this user or of similar users) in the physical world. The latter approach, which is the focus of this workshop, represents an emerging research strand whereby user models are augmented with real world knowledge to enhance adaptation and personalization.

Digital traces can be attributed to more than one individual, e.g. a circle of friends, a scientific community or even a whole population can be characterized by topics they tweet about, or things they comment about. Furthermore, events, e.g. conferences, local or global disasters, political debates, can be modeled by the streams of digital traces generated around these events (e.g. pictures, comments, discussions and reactions). Technological advancements, such as data/text mining, information extraction, opinion mining, social signal processing, interactive story telling, intelligent media annotation, semantic alignment, media aggregation and retrieval, make it now possible to automate the processing of digital traces to enrich system’s understanding about users’ experiences in the physical world. This technological development brings new opportunities to the user modeling community, and at the same time, opens up new technological, social, and ethical challenges.

The AUM workshop aims to create a forum for academic and industrial researchers and practitioners to discuss augmented user modeling from three angles:

- **Modeling**: methods and techniques for analyzing digital traces to capture, represent and connect user experiences:
  - What sources of digital traces can be used and are there any social and ethical constraints?
AUM Workshop Proceedings

- What aspects of user experience are captured in different digital traces and what techniques can be used to analyze digital traces?
- How can digital traces be processed, connected, and aggregated to provide useful information for modeling users and real-world activities/events?
- To what extent do these models represent what people, groups and events ‘really’ are in the physical world, e.g. do they conform to models and theories from social science?

- **Alignment:** methods and techniques for augmenting user models by aligning digital and real-world experiences:
  - How can digital traces be connected and represented to augment existing user models and to create holistic models of users, events, objects and groups?
  - How can parts of these holistic models be identified that are relevant to a certain user context - physical or digital?
  - How can different perspectives on activities and events be catered for and should they be aggregated in augmented user models?

- **Application:** personalization and adaptation approaches and application areas which can benefit from augmented user models:
  - How can adaptation and personalization approaches benefit from augmented user models?
  - What are the potential application domains (e.g. adaptive simulators, personalized virtual museums, personalized media retrieval, personalized information portals, personal assistants) and how can augmented user modeling improve the user experience in these domains?
  - Which types of personalization, recommendation and information filtering are possible and desirable for different applications or different real-world events (e.g. entertainment activities, job tasks, breaking news)?

We thank the members of the Program Committee of AUM 2011 for their support and reviews. Furthermore, we are grateful to all authors who submitted articles to AUM and contributed with their works to the AUM workshop.

Fabian Abel  Vania Dimitrova
Eelco Herder  Geert-Jan Houben

*AUM Organizing Committee, June 2011*

**Program Committee:** Ahmad Ammari, Kalina Bontcheva, Federica Cena, Ilknur Celik, Owen Conlan, Pasquale De Meo, Claudia Hauff, Dominikus Heekmann, Laura Hollink, Judy Kay, Alfred Kobsa, Tobias Ley, Stefanie Lindstaedt, Dhaval Thakker, Irene Pollach, Vincent Wade, Michael Yudelson and Ingrid Zukerman
Semantically Enriched Machine Learning Approach to Filter YouTube Comments for Socially Augmented User Models

Ahmad Ammari, Vania Dimitrova and Dimoklis Despotakis
Semantically Enriched Machine Learning Approach to Filter YouTube Comments for Socially Augmented User Models

Ahmad Ammari, Vania Dimitrova, Dimoklis Despotakis
School of Computing, University of Leeds
{A.Ammari, V.G.Dimitrova, scdd}@leeds.ac.uk

Abstract. Social media are media for social interaction that allow creating and exchanging user-generated content. The massive social content can provide rich resources for deriving social profiles that can augment user models and improve adaptation in traditional applications. However, potentially valuable social contributions can be buried within highly noisy content that is irrelevant or spam. This paper sketches a research roadmap toward augmenting user models with key user characteristics derived from social content. It then focuses on the first step: identifying and filtering noisy content to create data corpus about a specific activity. A novel, semantically enriched machine learning approach to filter the noisy content from social media is described. This is applied to a specific social source and activity: public comments on YouTube job interview videos. A potential application of the approach to augment user models in simulated experiential learning environments is discussed.

1. Introduction

The Social Web, or Social Media, includes a range of public data sources that are becoming an inevitable part in our life. Since their introduction, social media sharing sites such as YouTube\(^1\), Flickr\(^2\), and delicious\(^3\) have attracted millions of users, many of whom have integrated these sites into their daily practices. An inspection of the social video sharing platform YouTube reveals a high amount of community feedback through user comments on the published videos. These comments often include ‘authentic stories’ of people’s experiences of a particular activity. Pre-processing and mining these comments could provide a highly rich resource of real world activity descriptions based on individual’s and societies’ cognitive and social states, such as interests, knowledge, and experiences within that activity domain [14]. These identified features can be further mined to discover correlations between them that could then be used to augment existing and limited user models used to adapt many applications.

However, an important research challenge is how viable it is to extract the relevant content from within the huge amount of social media data that is likely to contain noisy content (i.e. content irrelevant to the activity of interest). The broad objective of our research is to evaluate whether social media content that is relevant to an activity

\(^1\) http://www.youtube.com/
\(^2\) http://www.flickr.com/
\(^3\) http://www.delicious.com/
of interest can be identified, mined, and used as an efficient source to augment user models used to adapt simulated learning environments.

The rest of the paper will include the following: In Section 2, we present a research roadmap towards achieving our broad objective. In Section 3, we describe a novel methodology to filter the noisy content from the social media data that we use to achieve our objective, which is the user comments on videos found on YouTube that describe a particular activity of interest. In Section 4, we position our work in the relevant literature on finding good quality content on the social Web by filtering the noisy content. In Section 5, we present and discuss the experimental results of our preliminary implementation. Finally, in Section 6, we discuss various considerations for subsequent implementations.

2. Socially Augmented User Models: Research Roadmap

Existing simulated learning environments suffer from the limited understanding of the learner because they are disconnected from the learners’ real job experiences. This often hinders learners’ engagement and motivation to undertake training since the skills developed in the simulated learning environment are not effectively connected to the skills used in the real job practice. Augmented User Modelling; i.e. enriching existing user models with additional information mined from other data sources not considered previously, is perceived as an approach to effectively help in aligning the learning experience in the simulated environments with the real world context and the day-to-day job practice. The key advantage is that the user models become aware of a range of aspects that cannot be captured from merely analysing the user interaction with the learning application.

Toward achieving the user model augmentation, we introduce a research roadmap, describing the research phases and the key research challenges that will be addressed.

Phase 1: Identifying social media content that represents real world user experiences. The key research challenge in this phase is how to filter the noise from the data sets retrieved from a given social media data source. By noise we mean those instances in the data sets that are highly irrelevant to a particular activity domain, thus not valuable for deriving significant features that can be used to augment existing user models with real world learning experiences.

Phase 2: Deriving key user characteristics from the clean relevant social content identified in phase 1. The key research challenge in this phase is how to derive social user profiles from the identified relevant content.

Phase 3: Using the social user profiles derived in phase 2 to augment an existing limited user model used to adapt a simulated learning environment. The key challenge in this phase is how to align the user in the existing user model with the social user profiles derived from the relevant social media content.

This paper focuses on the first phase in the roadmap. It presents a novel approach to filter the noise identified in the social media data. This hybrid approach combines machine learning, data mining, and semantics to address the challenge of this phase, which is the extraction of social media content that is highly relevant to a given real world activity of interest. The problem is narrowed down by considering a specific activity that is being practiced in the simulated environments. We use Job Interviews as the target activity, which is represented by videos selected from the social video
sharing site YouTube. The user comments found on these videos represent the corpus that will be processed by the approach to reduce the noisy content by filtering out those comments that are irrelevant to the particular activity domain of interest.

3. The Social Noise Filtering Approach

3.1 Filtering Noisy YouTube Comments: Methodology

In order to achieve a significant improvement in the relevance degree of the YouTube comments that are sufficiently good to derive key user characteristics for user model augmentation, we present a semantically enriched machine learning noise filtering approach. Figure 1 shows a flowchart representing the methodology for the approach.

**Step 1.** Select video corpus from YouTube about job Interviews. This was conducted as part of a research study to extract individual viewpoints from user comments in social spaces [4]. To illustrate the job interview activity, videos published on YouTube were selected as content source, and a thorough search and classification of different video types was performed. In particular, four different category types were identified to classify each retrieved video including: guides (explanations of best practices), interviewees’ stories, interviewers’ stories and interview mock examples. It was decided to focus on examples, as these resources can be closely connected to real world context representing the activity.

**Step 2.** For each selected video, retrieve the public comments on the video from YouTube. We call this Comment Collection A. Because this collection is retrieved from a very crowded and open social media sharing site, it contains a considerable rate of noisy comments. By noisy, we mean those comments whose text content is highly irrelevant (e.g. spam, abuse, etc) to the activity illustrated by the videos.

**Step 3.** Pre-process the Comment Collection A to build a Comment-Term Matrix (CTM) to train a supervised classification model. The goal is to represent each comment in the collection by a comment term vector. The pre-processing step is described in Section (3.2).

**Step 4.** Use the experimentally-controlled, relatively clean collection of YouTube comments collected and analyzed by the research study described in [4]. By clean, we mean comments whose text content is highly relevant to the job interview activity. We call this Comment Collection B.
Step 5. Analyze the Comment Collection B to build a **semantically enriched Bag of Words** (BoW). The resulted BoW forms a *ground truth vocabulary* that is highly relevant to the job interview activity domain. The selection and pre-processing of this comment collection are further described in Section (3.3).

Step 6. For each comment in Comment Collection A, **compute a relevance score** for the comment. Using the scores of the comments, label a new class attribute, i.e. a binary class attribute, with the distinct values: *relevant, noisy*, to supervise the learning of the classification model. This is further described in Section (3.4).

Step 7. Using the labelled Comment-Term matrix, train a **supervised classification model** that will learn the underlying classification rules to predict the *relevance, i.e. relevant, noisy*, of each new comment retrieved from the same data source, i.e. YouTube in the current case study, thus filter out those noisy comments deriving little-to-no key user characteristics for social user profiling.

3.2 Pre-Processing the YouTube Comments

Pre-processing the Comment Collection A is necessary to transform the textual corpus into a Comment Term Matrix (CTM) to be used as input data set to train classification models. A thorough description of the text pre-processing techniques to build Document – Term matrices to train machine learning models is found in [5]. The pre-processing steps to build the CTM are summarized in the following steps:

1. Remove all non-content bearing *stop words* like “a”, “an”, “the”, etc, which should not contribute to neither the representation of the comment nor to the scoring mechanism of each comment. A standard stop word list by Google⁴ has been used by this study.
2. Stem the words to retain the roots and discard common endings. The Iterated Lovins Stemmer [13] has been used widely for stemming unstructured data for machine learning and is therefore used by this study.
3. Rank the words based on their *tfidf* scores [1]. The *tfidf* score consists of two parts: term frequency *tf*, and inverse document frequency *idf*. A *tfidf* score is normalized between “0” and “1”.
4. Represent each comment by a Comment Term Vector, forming a Comment Term Matrix (CTM) representation of the Comment Collection. Each row in the matrix is a comment and each column represents a term and the value is the term *tfidf* score for that particular comment.

3.3 Building the Semantically Enriched Vocabulary

A clean, semantically enriched vocabulary / Bag of Words (BoW) that well represent the context of the job interview activity domain is needed to score each comment in Collection A. For this, we parse part of the corpus of study described in [4]. In that study, the selected YouTube videos were used in a system developed within the research context, and a research study is being conducted to collect video comments from participating users. The usage scenario for each participant includes: watching

---

⁴ [http://www.ranks.nl/resources/stopwords.html](http://www.ranks.nl/resources/stopwords.html)
the video; identifying useful video snippets; writing free text comments for each snippet indicating whether the comment corresponds to the activity presented in the video or a personal experience/opinion, and whether the comment concerns the interviewer or the interviewee. These comments provide examples of good (focused) corpus collected in experimental settings.

Figure 2 illustrates the corpus analysis phase. Each comment was handled as a separate document. The first step includes NLP techniques for text analysis using the Antelope NLP framework, i.e. sentence splitting, tokenization, Part of Speech tagging and syntactic parsing using the Stanford parser for linguistic analysis. This enables the extraction of a structured form text representation to empower further analysis using semantics. The second step consists of the semantic analysis layer, representing Ontology based word sense disambiguation and linguistic semantic text expansion. The first filter applied concerns the selection of specific lexical categories implemented within the WordNet Lexicon English language thesaurus to directly exclude non-significant terms for the job interview activity. For the words remained, the Suggested Upper Merged Ontology (SUMO) [3] has been exploited, which provides direct mappings of WordNet English word units to concepts in the ontology. The resulted concepts were used as word sense disambiguation indicators (second filter). In this context, WordNet Lexicon queries were performed to retrieve synonyms, antonyms and word lexical derivations to expand the word set. Furthermore, DISCO [8] has been exploited to retrieve distributionally similar words from the Wikipedia corpus, and the filters discussed above have been applied, i.e. lexical category and SUMO concept mapping.

3.4 Computing the Relevance Scores and Labelling the Comments

We present a mathematical model, using the Comment Collection A and the derived BoW in Section (3.3), to compute a numerical score for each public comment in
collection $A$, which represents the relevance of the comment to the job interview activity domain. Let $C$ be the set of all $n$ comments in the YouTube public comment collection $A$. For each comment $c_x \in \{ c_1, c_2, \ldots, c_n \}$, there is a set $w_{cx}$ of unique tokenized and stemmed non-stopwords, where $m$ is the number of these words in comment $c_x$. Let $B$ be the set of all the stemmed and unique words in the BoW derived in Section 4.3. We then define a relevance score $S_{cx}$ for the comment $c_x$ to be:

$S_{cx} = \frac{|w_{cx} \cap B|}{(\sum_{k=1}^{n}|w_{ck} \cap B|) / n}$

where $|w_{cx} \cap B|$ is the number of words that exist in the intersection between the sets $w_{cx}$ and $B$, and the denominator is the average number of words that exist in the intersections between each set $w_{ck}$ and $B$, where $k \in \{1, 2, \ldots, n\}$.

In order to train a binary classification model, we define a target class attribute $\text{CLASS}_{cx}$, which contains a nominal value $\in \{ \text{noisy} \ (0), \ \text{relevant} \ (1) \ }$, based on the value of the score $S_{cx}$ for the comment $c_x$:

$\text{CLASS}_{cx} = \begin{cases} 
\text{noisy} \ (0) & \text{if } S_{cx} < 1.00 \\
\text{relevant} \ (1) & \text{if } S_{cx} \geq 1.00 
\end{cases}$

The class value for each comment is then assigned as the target class attribute value to the term vector representation of the comment, forming a supervised training corpus for building machine learning classification models that learn the underlying classification rules to predict the class value of new comments.

<table>
<thead>
<tr>
<th>Comment Labelled Noisy</th>
<th>Score</th>
<th>Comment Labelled Relevant</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>what if you never had a job</td>
<td>0.34</td>
<td>To be honest, I probably wouldn't hire either one of them. The girl is obvious, but the guy's leg twitching bothered me, as did his leaning forward in the chair, and he focused too much on his past. I want to hear what he's going to do with the job available, not so much what he has done.</td>
<td>5.08</td>
</tr>
<tr>
<td>LOL</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interview on wednesday hope it goes well</td>
<td>0.68</td>
<td>that part when she answers her phone was just retarded, AHHHHHH! someone's calling me! the person giving the interview must think she's psychopathic</td>
<td></td>
</tr>
<tr>
<td>come see my job interview come see my job interview come see my job interview called Boss Boss Baby Boss Boss Baby</td>
<td>0.79</td>
<td></td>
<td>1.13</td>
</tr>
</tbody>
</table>
To give a sense of the reasonability of the scores and labels assigned to the comments based on our model, table 1 shows four example comments on the left that have been labelled as noisy by the scoring model. Obviously, the first three ones do not comment on the job interview video being watched, whereas the fourth one is a spam. The scoring mechanism was reasonable in labelling them as noise even while containing a considerable number of words, i.e. comment 4. The two comments on the right clearly describe actions occurring within the activity watched in the video, thus potentially can derive user characteristics related to the activity. Again, it was reasonable labelling them as relevant.

4. Related Work

There have been a few attempts in the literature to create information filtering mechanisms for adaptation in the social Web, which can be linked to the research challenge addressed in our study. For example, the work in [11] presents CompleXys, a system that accesses a variety of social data sources, including social networks and blogs, and semantically annotates and categorizes the retrieved content based on a filtering layer and displays only the relevant content to the user. The filtering layer takes the output of a content annotator component that annotates the retrieved content using a domain ontology. The expanded taxonomy is then meant to decide whether a given resource is relevant to the list of topics stored in the filtering layer. The frequency of occurring annotations can then be used as a simple indicator for the relevance of a certain topic. We have further expanded this mechanism by introducing the mathematical model in Section (3.4), which computes a relevance score for each retrieved content observation, i.e. YouTube video comment, and then labels the observation, i.e. relevant or noisy, accordingly.

Works on filtering spam blogs (or splogs) [15] as well as filtering blog spam comments [6] could also be linked to this study. In [15], blogs and their connections are represented as a graph and then various graph statistics, i.e. degree distribution, clustering coefficient, are computed. It is shown that these statistics are considerably different between splogs and legitimate blogs, and therefore could be leveraged to identify splogs. The work in [6] presents a similar approach to identifying spam comments irrelevant to the discussion by generating a blogger network based on the blogger’s commenting behaviour. However, social comments in general contain no (or very little) hyperlinks between them. This leads to a highly sparse adjacency matrix with very few non-zero values that represent the link strength between the comments [1]. Computing content-based similarities between the comments could be used to fill the matrix in addition to the direct links to reduce sparsity. However, since comments usually do not contain much text, content-based estimation of the comment linkage is not a good alternative and the underlying noise filtering approach is likely to perform poorly in noisy comments identification.

Few works have used machine learning to find quality contents from the user comments on the social space. The work in [2] used binary classification models to automatically identify high quality content in a large community-driven question/answering portal; Yahoo! Answers. We further extend this work by introducing semantic enrichment in Section (3.3) in order to classify the data set used for training the binary classification models. The work in [12] used a supervised
A classification approach to analyze a corpus of YouTube comments in order to discover correlations between the user views and sentiments extracted from these comments, and the comment ratings by the readers of these comments. Such correlations may help to automatically structure and filter comments for users who show malicious behaviour such as spammers and trolls. However, relying on a comment rating needs a huge corpus of these comments because just a small fraction of the comments on YouTube is rated by the YouTube community. This large size of corpus is not always available when addressing a particular domain activity. For example, a total of 17 high quality YouTube videos on the “Job Interview” activity selected for the work of this paper did not retrieve more than 1159 comments. Instead of relying on comment ratings, the approach presented in our work creates a semantically enriched taxonomy by analyzing a clean corpus of experimentally-controlled user comments and enriching this vocabulary with semantic annotations to form a ground truth Bag of Words (BoW) that is highly relevant to the activity domain of interest, i.e. job interview. The retrieved YouTube comment corpus is then scored and labelled, using the mathematical model and the semantically-enriched BoW, and then used to train a supervised classification model that predicts and filters out the noisy comments.

5. Experimental Results

A preliminary implementation to the approach has been done to evaluate the classification performance in filtering the noisy comments from the training / testing corpus. Table 2 shows a summary description of the two comment collections before and after being pre-processed.

<table>
<thead>
<tr>
<th></th>
<th>Comment Collection A</th>
<th>Comment Collection B</th>
</tr>
</thead>
<tbody>
<tr>
<td>No of Videos</td>
<td>17</td>
<td>5</td>
</tr>
<tr>
<td>No of Comments</td>
<td>1159</td>
<td>193</td>
</tr>
<tr>
<td>Min Intersection Size</td>
<td>0.0</td>
<td>6398</td>
</tr>
<tr>
<td>Max Intersection Size</td>
<td>48.0</td>
<td>25606</td>
</tr>
<tr>
<td>Avg Intersection Size</td>
<td>8.85</td>
<td>1978</td>
</tr>
<tr>
<td>Min Relevance Score</td>
<td>0.0</td>
<td>17604</td>
</tr>
<tr>
<td>Max Relevance Score</td>
<td>5.55</td>
<td>79204</td>
</tr>
<tr>
<td>Avg Relevance Score</td>
<td>1.00</td>
<td>Total after Stemming &amp; Removing Duplicates 4382</td>
</tr>
</tbody>
</table>

17 YouTube videos have been selected to retrieve 1159 comments for collection A. Five of these videos have been used so far to collect 193 user-guided comments for collection B. Analyzing these comments has derived 4382 unique words relevant to the job interview activity, forming our semantically enriched BoW. For the trial of this paper, we chose to expand the original words of the comments with synonyms,
antonyms, derivations, and DISCO entries. In future implementations, we aim to further expand the vocabulary with the remaining resources as described in Section (3.3). Applying the relevance scoring and labelling model described in Section (3.4) on collection A comments have assigned 724 comments as noisy and 435 comments as relevant. Text pre-processing these comments has derived a CTM matrix having 1159 comment term vectors and 903 predictor attributes representing the \( tfidf \) term weights, in addition to the target binary attribute containing the class value (noisy or relevant) of each training comment.

We have used the labelled CTM as a training corpus to train two types of classifiers widely used for document classification, C4.5 Decision Tree [9] and Naïve Bayes Multinomial [7], to evaluate predicting noisy comments that should be filtered out when retrieving further YouTube comments to be used for deriving key user characteristics directly relevant to the job interview activity. The C4.5 algorithm has the ability to auto-detect those predictors most contributing to the target class and use them in the underlying classification rules. Naïve Bayes Multinomial (NBM), on the other hand, is a probabilistic classifier that has achieved good prediction results in spam filtering [10]. We used three different training / testing corpus variations to train three models from each classifier to test the prediction stability performances. In the first variation, we test the classifiers on the same full dataset the classifiers are trained on, whereas in the second and third variations, we trained the classifiers on 80% and 60% of the full dataset, respectively, and tested on the remaining instances.

<table>
<thead>
<tr>
<th>Testing Size</th>
<th>C4.5 Full</th>
<th>C4.5 80%</th>
<th>C4.5 60%</th>
<th>NBM Full</th>
<th>NBM 80%</th>
<th>NBM 60%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correctly Classified Comments</td>
<td>1070 (92.3%)</td>
<td>194 (83.6%)</td>
<td>390 (84.1%)</td>
<td>1063 (91.7%)</td>
<td>189 (81.5%)</td>
<td>362 (78.0%)</td>
</tr>
<tr>
<td>MAE</td>
<td>0.14</td>
<td>0.22</td>
<td>0.22</td>
<td>0.10</td>
<td>0.21</td>
<td>0.24</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.26</td>
<td>0.38</td>
<td>0.37</td>
<td>0.27</td>
<td>0.41</td>
<td>0.45</td>
</tr>
<tr>
<td>TP Rate</td>
<td>0.98</td>
<td>0.90</td>
<td>0.92</td>
<td>0.92</td>
<td>0.85</td>
<td>0.76</td>
</tr>
<tr>
<td>FP Rate</td>
<td>0.16</td>
<td>0.28</td>
<td>0.31</td>
<td>0.09</td>
<td>0.24</td>
<td>0.18</td>
</tr>
<tr>
<td>Precision</td>
<td>0.91</td>
<td>0.85</td>
<td>0.85</td>
<td>0.95</td>
<td>0.86</td>
<td>0.89</td>
</tr>
<tr>
<td>Recall</td>
<td>0.97</td>
<td>0.9</td>
<td>0.92</td>
<td>0.92</td>
<td>0.85</td>
<td>0.76</td>
</tr>
<tr>
<td>ROC Area</td>
<td>0.93</td>
<td>0.84</td>
<td>0.82</td>
<td>0.97</td>
<td>0.87</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Table 3 shows the evaluation metrics for the six trained models. The average correctly classified comments by the C4.5 algorithm is 86.7%, slightly higher than for the NBM algorithm, 83.7%, resulting in a slightly lower Root Mean Squared Error (RMSE) for C4.5 (0.34) than it is for NBM (0.38). However, the average Mean Absolute Error (MAE) for C4.5 and NBM are almost the same, 0.19 and 0.18, respectively. The True Positive (TP) rate is the rate of correctly classified noisy comments to the total number of noisy comments in the testing dataset. On average,
C4.5 is more able to correctly classify noisy comments from within the total available noise than NBM. However, NBM is less likely than C4.5 to misclassify relevant comments that may derive important user characteristics as noise from within the total relevant comments available. This is noticed in the lower False Positive (FP) rate for NBM than it is for C4.5, as well as for the higher Precision rates for NBM.

The Classifier Output also gives the ROC area, which reflects the true positive rate versus the false positive rate. This metric reflects the probability that a randomly chosen noisy comment in the testing data is ranked above a randomly chosen relevant comment, based on the ranking produced by the classifier. The best outcome is that all noisy comments are ranked above all relevant comments, in which case the ROC is 1. In the worst case it is 0. Figure 3 depicts the ROC curves for C4.5 (a) and NBM (b) both tested by the full data set (n = 1159), with FP Rate on the x-axis and TP Rate on the y-axis. NBM shows a slightly larger ROC area (0.90) than C4.5 (0.86). Moreover, NBM needs less costly misclassifications of noise (FP rate) than C4.5 to reach the optimal desired correct predictions of noisy comments (TP rate).

In general, the output of the experimental study – the classification evaluation metrics – shows that the two classifiers implemented provide good performance in predicting and filtering out the noisy YouTube comments that are irrelevant to the particular activity domain of interest (job interviews). Although the C4.5 decision tree classifier is slightly better in filtering the noisy comments from the total available noise, the Naive Bayes Multinomial classifier shows less risk in misclassifying relevant comments, which can derive key user characteristics to augment user models, as noise. In addition, the comment relevance scoring and labelling model proposed in Section (3.4) provides a reasonable estimate to whether each comment within the classification training corpus could be considered either noisy or relevant to the sought domain activity, i.e. job interview.

As discussed in Section 2, filtering out the irrelevant content from the noisy social media data is considered as the first phase in our research roadmap toward utilizing social media content to augment existing user models. After removing the identified noise, the remaining relevant YouTube comments will then be used to retrieve additional YouTube content generated by the users who posted these comments. These may include meta-data about any videos that these users upload or mark as favourites, additional comments they may post on YouTube, and explicit information that the users may write about themselves on their YouTube profiles. All these user-
generated contents will then be analyzed further to derive the social user profiles for those YouTube users. These profiles will then be mined to discover interesting associations between the several user characteristics that these profiles consist of. Finally, the revealed associations will be exploited to augment existing user models for similar users who use simulated learning environments for experiential learning.

6. Future Work

For future implementations of the approach, we aim to take several considerations into account to further improve filtering results, as summarized below:

- Further statistical analysis of the comments in the training corpus (collection A) could be conducted, in order to improve the accuracy of the scoring mathematical model. Comparisons with other variations, such as considering the comment size rather than and in addition to the comment intersection with the ground truth bag of words are also aimed. Expert-based evaluation of the computed scores and labels are also important to reduce false learning of the classification rules by the trained classifiers.

- Further semantic enrichment to the ground truth vocabulary by considering the ontologies described in Section (3.3) will be conducted. Weighting the original words derived from the controlled comments as well as the semantic expansions to these words by their importance to the activity domain of interest is also aimed to improve the accuracy of the relevance scoring mechanism.

- Further evaluations and comparisons with more classifiers that provide good classification results with unstructured data. These include variations to the Naïve Bayesian algorithm, Singular Value Decomposition-based algorithm, Support Vector Machines, and Combination algorithms. Moreover, classifier-specific parameter tuning and dimensionality reduction to the training comment-term matrix will be applied to further improve the prediction accuracy.

Acknowledgement

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no ICT 257831 (ImREAL project).

References


7 http://www.imreal-project.eu/
The Personal Adaptive In-Car HMI: Integration of External Applications for Personalized Use

Sandro Rodriguez Garzon and Mark Poguntke
The Personal Adaptive In-Car HMI: Integration of External Applications for Personalized Use

Sandro Rodriguez Garzon¹ and Mark Poguntke²

¹ Daimler Center for Automotive Information Technology Innovations, Berlin, Germany
sandro.rodriguez.garzon@dcaiti.com
² Daimler AG, Ulm, Germany
mark.poguntke@daimler.com

Abstract. We describe an approach for integrating non-automotive applications into in-car-entertainment systems while taking account of manifold personalization capabilities within a mobile environment. Adaptive user interfaces are generated for external applications using well-known interaction and personalization concepts. The interaction concepts are defined via state-based interaction models and utilized for the integration of various applications in order to guarantee a common look and feel. Context-aware adaptations of the user interfaces are achieved by supporting the process of gathering an augmented user model with a personalization concept in form of personalization guidelines. We present and discuss an exemplary application for a personalized, safe in-car HMI that automatically adapts to the targeted design and interaction concept as well as to the personal needs of the user.

Keywords: personalization, adaptation, user modeling, user interfaces, interaction modeling, model-based development, automotive apps

1 Introduction

The American Dialect Society has chosen "App" to be the word of the year 2010 [18]. Apps are mobile applications that can be downloaded and installed on high-performance smartphones. These apps are more and more used for daily tasks and often use content from online sources. The individual extent of usage, however, strongly depends on the individual user. One user may need only one or two apps helping him getting the latest news or the weather forecast tomorrow. Another user may use hundreds of applications and organize important parts of their daily life around smartphone apps and web applications. What all groups of users have in common is the predictability of their behaviour in similar situations [4, 16]. They may use single functions of the same application again and again depending on their respective personal context of use.

In an in-car environment it is required to have minimized driver distraction in the human-machine interface (HMI) for infotainment applications. Furthermore,
the use of mobile devices while driving is prohibited by law in most countries. A need for seamless integration of external applications to the in-car environment arises considering the amount of possible apps a driver may want to use in their daily life.

We propose a system supporting the integration of external applications to the in-car HMI and supporting the predictability of user behaviour by adapting the HMI according to different repetitive situations. We propose to use abstract interaction models and extend these with elements for personalization. This allows to use the same abstract model as a basis for different design and interaction concepts.

2 Related Work

The above presented situation implies two aspects of user interface engineering. An approach for user interface generation is needed. In addition, the process has to support user interface personalization.

2.1 User Interface Generation

Generating user interfaces from abstract interaction representations requires an approach for abstract interaction modeling and respective transformation processes. Model-based development using a respective modeling language holds the advantages of reusability, readability and easier adaptability of user interfaces amongst others [9][7]. A common basis to start with user interface specification is a task model. Concur Task Trees (CTT) [11] provides a notation for task models that can be used to derive user interfaces in subsequent steps. The User Interface eXtensible Markup Language (UsiXML) [15] describes a modeling and transformation approach from abstract to concrete user interfaces based on the Cameleon reference framework [1]. In recent years, several approaches motivate the use of the Universal Modeling Language (UML) [10] for user interface modeling. UML has been used for software engineering processes for many years with different established modeling tools. Also, several proprietary tools for model-based development exist for user interface specification, simulation and the automated generation of code or final user interface descriptions [2] [3] [15]. However, the generation of a respective final user interface from abstract models at runtime is not covered by these approaches since they focus on design time aspects.

2.2 User Interface Personalization

The adaptation of user interfaces based on the analysis of its historical use is applied in a lot of application areas. Inspired by Mark Weiser’s vision on ubiquitous computing [17], Ma [8] introduces a smart home environment that observes the resident behavior in order to automate common tasks. Using a case-based reasoning approach the smart home environment is able to adjust a TV channel
or the air conditioning depending on the residents preference within a certain context (e.g. time). Coutand [5] presents a Call Profile Service that adapts its behavior based on a detected user preference at a certain location. It observes the way the mobile phone user accepts incoming calls and adjusts the profile accordingly. A similar mobile phone application for automatic profile selection was described by Schmidt [14]. Schmidt furthermore uses the knowledge about the user’s activity (walking or stationary) and the light conditions to adjust the font size within a notepad application. The difference to the former approaches lies in the fact that Schmidt’s solution does not consider the user specific preference. Predefined rules are utilized to detect the current context and to adapt a device in a "one-fits-all" manner. The results of a user study presented in [4] indicate that users like to arrange the icons of their mobile phone applications by relevance and the current activity. Since relevance is a user specific preference, the prototype needs to log every rearrangement of each user and the corresponding context in order to propose user-centric and context-dependent icon arrangements.

3 Personalization of Generated User Interfaces: Use Case

In this paper we focus on integrating external applications or services into the in-car head unit while providing the possibility to personalize. We introduce the mobile application called Supermarket Guide. Its purpose is to provide information about nearby supermarkets and their current offers. This external application can be integrated to the in-car head unit. It may come from an external device or from an online source. The application will appear as an additional entry within the application line next to preinstalled applications like Navi and Phone. Figure 1 shows screenshots of the supermarket guide that was automatically integrated to the head unit based on an abstract model and a transformation process. The main view appears underneath the application line as illustrated in Figure 1. For each entry within the supermarket list, the user is able to view either how to approach a supermarket or a list of supermarket-specific offers.

In order to describe the personalization that happens within the HMI of the Supermarket Guide, we will introduce a sample usage situation that is illustrated in Figure 2. Consider a user that lives in the city of Stuttgart and commutes every workday to the city of Ulm. After work, the user always drives back towards his home in Stuttgart. Sometimes - in case the user needs to buy food - the user has a stopover at one of his preferred supermarkets along the route: supermarket A in Ulm or supermarket B in Stuttgart. Immediately before approaching junction M, the user starts the Supermarket Guide and checks for interesting offers of supermarket A in order to decide whether or not to make a stopover in Ulm. If the user isn’t satisfied with the offers he continues and checks the offers of supermarket B before approaching junction N. Otherwise, the user will stop at supermarket A.
After repeating the same interactions regularly, a light bulb symbol will appear next to the icon of the supermarket application each time the car approaches the road before junction N or the road before junction M. The light bulb indicates that there is a personalized adaptation of the interaction. If the user starts the Supermarket Guide while the light bulb is visible, the supermarket application will behave differently. The application will not open the main view but a supermarket-specific offers view: either the offers of supermarket A while driving on the road next to junction M or the offers of supermarket B while driving on the road next to junction N. Generally spoken, the icon signalizes that the HMI of the Supermarket Guide has a user-specific behavior as long as the light bulb remains visible. That is, the HMI adapts its behavior according to the users needs be means of observing real world experiences.

4 Abstract Modeling for User Interface Generation

We use the roles of an application developer and an interaction designer for the abstract modeling approach. An application is developed by an application developer including a functional application interface consisting of a class diagram with attributes and operations. An interaction designer uses this interface to create an abstract interaction model using UML state charts to describe user actions and corresponding system reactions. A transformation program uses the model and generates a user interface compliant to the respective automotive HMI concept. For the transformation process rules have to be implemented mapping the abstract model elements to user interface elements for a specific concept. [12]
Considering the *Supermarket Guide* an abstract interaction model consists of a list representation with two selectable options for each list entry. An abstract model is illustrated in 3. The *List Overview* state contains the presentation of the supermarket entries. *Entry Detail 1* and *Entry Detail 2* present the respective information depending on the selected option, *Offers* or *Directions*, for the selected entry. A class diagram including required variables and operations as well as the interaction model using these defined elements has to be provided along with the application to be integrated which is in the given use case the *Supermarket Guide* application. Since the focus of this paper is the general approach to extend abstract interaction models with personalization adaptions we use only the illustration of a general model as in 3. Details on the modeling approach can be found in previous work [12].
5 User Interface Personalization

In order to support user interface personalizations as described in chapter 3 we introduce a real-time component that has to deal with the process of detecting regular user behavior within similar environments. Thus, the main challenge of this so called personalization component lies in observing the real world user behavior and inferring the right moment to notify the HMI about an upcoming user preference. The correlation between the user preference and the real world user behavior will be stored within an augmented user model.

In case of the Supermarket Guide, the personalization component observes the way the user interacts with the application-specific HMI at different locations. If the personalization component detects a user-specific regularity within a certain region it will notify the HMI about the user preference every time the user enters the region. This information can be processed by the interface in order to provide personalized functionality.

5.1 Personalization Component: In- and Output

Considering the personalization component of the Supermarket Guide as a black box, it is necessary to notify the black box whenever the user opens the supermarket offers view. This information gives the black box a first glance onto the amount of times the user likes to view the supermarket offers. Enriching the notification with the supermarket-specific name allows the black box to detect which supermarket-specific offers are favored by the user. Additionally, the user prefers to query different supermarket-specific offers at different locations. Therefore, the black box also needs to get notified about the environment in which the user interaction is happening. Hence, different digital traces need to be observed by the black box in order to detect and output regular environment-specific user preferences.

The content of the different digital traces can be distinguished by its origin. The event of "opening a supermarket offers view" as well as the supermarket-specific name are declared to be application specific knowledge. In contrast, the environment-specific information - in our use case: the road - is declared to be a system-specific knowledge. This kind of knowledge is provided by the system independently of the type of applications that are preinstalled on the head unit.

All events originating from human-computer interactions are called interaction events regardless of its origin. The black box itself does not differentiate between events having different origins. Thus, the task of detecting regular situation-dependent user behavior is based solely on the processing of a single stream of interaction events. In other words, the personalization component provides two interfaces: An input interface for interaction events and an output interface for notifications about the occurrence of a situation that will likely contain a regular user interaction.
5.2 Personalization Component: Internals

Since every application has its own user interface with different personalization scenarios it is necessary to let an expert define the way the personalization component should feed its augmented user model. The expert, known as the personalization designer, configures the personalization component by means of personalization guidelines in form of a scenario specification file. Therein, each configuration step is closely coupled with a certain subprocess of the personalization component. In the following, every subprocess will be introduced briefly with its required configuration steps.

**Action Discovery** Firstly, each incoming interaction event will be passed to the action discovery subprocess. The purpose of this process is to detect sequences of interaction events that are supposed to be relevant concerning the specific scenario. The concrete occurrence of such a sequence is called action. Considering the Supermarket Guide, an action contains the event that is generated by the HMI in case the user enters the supermarket offers view. But such an action is only a valid detected action if the user remains at least 10 seconds at the supermarket offers view. This additional condition restricts the action discovery process to detect actions that are originally intended by the user and not executed by mistake. These conditions are defined by the personalization designer by means of a temporal event pattern that describes the relevant action in a general way. The result of the subprocess is an action consisting of a sequence of one or more concrete interaction events.

**Context Discovery** Each detected action will be received by the context discovery subprocess. This subprocess groups actions that happened within similar environments. Considering the Supermarket Guide, two actions are grouped in
case the car was driving on the same road at the moment the action was detected. A group is declared as significant if it contains at least a minimum number of actions. Only significant groups are contributing to the augmented user model as valid regularities. This becomes necessary in order to ignore user interactions that are executed rarely. The subprocess is guided by the personalization designer through the specification of the environmental factors which are relevant to the scenario.

**Situation Discovery** The last step - *situation discovery* - deals with processing each significant group in order to detect similar situations. Considering the Supermarket Guide, the user expects the HMI to personalize itself in case he opened the offers view of a specific supermarket several times at the same location. Since the input stream of interaction events also contains location events it is straightforward to consider the stream of interaction events in conjunction with the regularity found in the previous subprocess. The personalization designer decides by means of a temporal event pattern how both information streams are joined to detect the desired situation. For the Supermarket Guide, the interweavement of both streams is fairly simple: Look for location events with a road similar to the road of a significant group. If a situation is found the personalization component will notify the HMI. The reader is referred to [13] for a more detailed example.

### 5.3 Extension of the Interaction Model

The HMI in turn is in charge of providing the application-specific interaction events and to execute a personalization in case it is notified about a similar situation. Since the HMI is generated based on an interaction model, it is necessary to extend the interaction model in order to deal with the personalization task.

Considering the HMI of the Supermarket Guide, two interaction events needs to be generated: "Opening a supermarket offers view" and "Leaving a supermarket offers view". Since we have already specified a state for the supermarket-specific offers view it is straightforward to define entry and exit actions to fire the required events. We assume that all states implicitly fire entry and exit events. Thus, it is not necessary to explicitly define these kind of actions.

As mentioned in Section 3, the personalization of the Supermarket Guide will be announced by the light bulb symbol appearing in the application line next to the shopping cart icon. This personalization icon is an internal feature of the system and is reused for every application that offers personalization capabilities. In order to jump directly into the supermarket offers view at the start of the application, it is necessary to add a conditional transition immediately after the entry point. Figure 5 illustrates the modified interaction model with the extension colored in blue. It forces the HMI to check whether the personalization component announced a personalization. If there is no announcement the HMI will show the main view. Otherwise, the HMI will show the supermarket offers view of the announced supermarket. A second conditional transition is added
in order for the interaction model to be reusable for other applications with different personalizations.

![Adapted State Diagram for Personalization](image)

**Fig. 5.** Adapted State Diagram for Personalization

### 6 Discussion

Up to this point, the interaction model of the Supermarket Guide was used to generate the application-specific HMI. Hence, the external application gets a well-known HMI concept rather than providing its own. The benefit of having an abstract interaction model becomes obvious in case a second application needs to be integrated. Consider the integration of the social networking platform [Facebook](https://www.facebook.com) using the same general interaction model. Figure 6 illustrates a proposed HMI concept for an in-car Facebook application. Since the main structure of the HMI with its main view and two detail views matches with the HMI of the Supermarket Guide, it makes sense to reuse the interaction model with a different parametrization. Substituting the data sources and modifying the specification for the personalization component are the only modifications that are necessary. The latter modification arises from the fact that the Facebook application will be personalized in a slightly different manner. The personalization decision to enable direct jumps into one of the detail views does not depend upon a location information provided by the car system but from the abstract location that is assumed to be provided by the Facebook service. Out of the perspective of the personalization component it does not matter if the interaction events are generated by a system component or by an external source. Even multiple event

---

3 Facebook is a trademark or registered trademark of Facebook, Inc.
sources like location and the kind of persons within the car can be used to decide whether to show the wall or info views.

Thus, the major benefit of the presented approach lies in the unobtrusive integration of external and adaptive applications by using well-known interaction concepts. New adaptive applications can be developed faster because the automotive-specific HMI concepts are known in advance. The developer isn’t in charge of dealing with the special issues of user interface development within the automotive environment like distraction reduction. Additionally, all potential adaptations of the HMI are also known in advance since they are an integral part of the interaction model. On the one hand, the user has a clear understanding about the adaptation that might happen. On the other hand, the application developer in conjunction with the personalization designer is still able to specify the generic conditions describing the moment an adaptation might occur.

It remains to be seen whether it is practicable for a personalization designer to specify all the personalization conditions in a generic form using a temporal event pattern as described in Subsection 5.2. In our approach, the interdependence between different digital traces must be known a priori to properly configure the subprocess of the personalization component. Sometimes it is not predictable which environmental factors are influencing a user’s intention of changing for example the radio station.

Another question arises concerning the way personalization is detected. In some cases it might be reasonable to integrate the personalization detection
within the application itself. For example, the automatic reordering of application-specific list entries depending on the user interest is an application-specific personalization. It does not make sense to share these kind of detected regularities with other applications and thus do not require a system-wide augmented user model. But, as stated above, one of the main benefits of user interface generation based on abstract models is the reuse of interaction models for different applications. Thus, extending the interaction models in order to handle personalizations might affect multiple applications. A system-wide personalization component would be able to detect a regularity in an application A and could provide this information to an application B. The user interface of application B would then adapt itself since it is originally based on the same interaction model. Even newly installed applications would instantly profit from regularities that were detected so far.

7 Conclusions and Future Work

We presented an approach to consider user-centric adaptation mechanisms within the definition of state-based interaction models. It was demonstrated that a properly specified interaction model together with a clearly defined personalization is sufficient to integrate non-automotive applications into the automotive environment. Particularly, the special requirements for the automotive environment enforced us to develop a guided personalization process in order to guarantee a predictable and comprehensible HMI behavior. Considering our approach, an application developer is required to select an adequate interaction model which is provided for example by a car manufacturer. The personalization remains still flexible and can be configured by the application developer together with the personalization designer via the configuration of the personalization component.

A working prototype for the HMI generation process [12] as well as a prototype for the personalization concept [13] were demonstrated previously. In future work, these implementations have to combined to further evaluate the presented approach.

References

10. OMG. Uml 2.2 superstructure specification, 2009.
A Semantic Approach to Extract Individual Viewpoints from User Comments on an Activity

Dimoklis Despotakis, Lydia Lau and Vania Dimitrova
A Semantic Approach to Extract Individual Viewpoints from User Comments on an Activity

Dimoklis Despotakis, Lydia Lau, Vania Dimitrova

University of Leeds, School of Computing, UK
(sccd, V.G.Dimitrova, L.M.S.Lau)@leeds.ac.uk

Abstract. A vast amount of user contributed content about real world human activity exists in social web spaces (e.g. videos and user stories or comments in YouTube) rich of personal experiences and opinions which intuitively reflects reality settings. This content can be a useful source for enriching the learning experience in simulated environments, if exploited in an appropriate context for the learner. An interesting observed characteristic of the user contributed content is the diversity of viewpoints. A novel approach for multi-viewpoint knowledge elicitation and representation to enable intelligent content retrieval is being investigated, and conducted within one of the use cases of the ImReal EU project. The job interview process has been selected for studying the intricacy in interpersonal communications, focusing on the emotion and body language signals dimensions, and the different individual viewpoints of the activity. Work in progress presented in this paper includes an analysis of the problem domain, the information extraction process and preliminary results.

Keywords: Activity, Individual viewpoint, Emotion, Body Language

1 Introduction

Simulated environments, where learners are involved in simulated situations and perform activities that resemble actual job activities provide powerful learning tools for developing soft skills in ill-defined domains[1]. Such environments will have a strong presence in the future intelligent learning technologies, especially in the area of workplace training, where adaptation and personalization will play a key role[2]. Conventional adaptation approaches model users based on their interaction within the simulated environment and adapt accordingly. However, to be effective, training environments for adults should offer learning experiences directly relevant to the real world job context and aligned with the learner’s needs in practice [3]. The challenge is that real world activities are affected by dynamic conditions and complex situations which are hard to capture into the simulated world, whereas a simulated environment embeds predefined scenarios with fixed parameters.

On the other hand, there is a vast amount of user contributed social content about real world activity (e.g. user comments or stories) providing rich content of information about people’s personal experiences and opinions. This abundance of
user generated digital content provides potentially useful traces that can reflect what happens in the real world. Although this content can be a useful source for enriching the learning experience and bridge the simulated settings with reality, it has not been exploited to date, designating a key research challenge:

*Can digital traces from the social spaces be used to construct a model of the real-world activity and context, and how can this model improve adaptation in simulated learning environments and enable intelligent content retrieval?*

As part of the above research challenge, in this paper we present work in progress on the implementation of a novel approach to collect and analyze user comments on social media, particularly comments on videos populated in a YouTube-like environment, in order to:

*Identify key concepts related to specific activity and determine individual viewpoints on the activity*

An ontology-based information extraction process is proposed to support the above key objectives. The Job Interview activity has been selected as a case study for this work. Section 2 presents the case study (focusing on emotions and body language in interviews), and provides an initial analysis of the potential use of the comments collected by a YouTube-like environment. Section 3 reports on the technicalities and the implemented information extraction algorithm, while Section 4 presents the results to date. Finally, Section 5 concludes with a discussion on the current approach with regard to related work, and future plans.

## 2 Case Study – The Job Interview Activity

This research is being conducted within one of the use cases of the ImReal EU project, which aims at developing simulated environments for Immersive Reflective Experience-based Adaptive Learning. The job interview process is selected for further investigation as it exemplifies a key challenge that ImReal addresses: developing soft skills within simulated environments for training in work-place interpersonal communication. Additionally, a wealth of related media content is available on the web, which can be used to illustrate our approach.

The focus of this study is on capturing users’ experience on the effect of emotions and body language which may affect the interpersonal communications. The videos act as a catalyst to simulate discussions and recall of personal experiences.

### 2.1 Content Collection

In order to enable users to make comments on specified snippets within an online video, a system has been developed to: (i) provide links to a sample of YouTube videos on job interview (Fig 1) and (ii) enable the participants in the research study to

1 http://www.imreal-project.eu
interact with a video by selecting snippets (Fig 1, a) and commenting on each snippet (Fig 1, b) (please refer to Section 4 for descriptive details). A snippet has a start and stop time relative to the beginning of the video.

![Fig. 1. A screenshot from the implemented snippet to collect content. Participants (a) partition the video into snippets and (b) add comments indicating the actor and if it relates directly to the video or it refers to the participant’s personal experience.](image)

The comments are in free text. Users can indicate (i) whether a comment corresponds to the interviewer or the interviewee, and (ii) whether it relates directly to the video (Table 1, C1) or it refers to the user’s personal life (e.g. an experience or opinion) which was triggered by observing this particular video snippet (Table 1, C2). Indicating the actor of the activity (i.e. the interviewee or the interviewer) that the comment refers to resulted after evaluating the elicitation process with Social Scientists, who showed that the actor of the activity is a core component for analyzing the comments. Furthermore, providing personal experience related comments allows the potential to capture not only important concepts as presented in a video but also key aspects that are possibly missing in the video and reveal relevance to real life.

The output of this YouTube-like environment is a collection of snippets with comments from users who have been watching the videos.

### Table 1. Comments of the same participant on a video snippet. C1 relates directly to the video, while C2 refers to the participant’s personal experience. Both correspond to the interviewee.

<table>
<thead>
<tr>
<th>Comment Text</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C1</strong>: “The interviewee rushes into the room.”</td>
</tr>
<tr>
<td><strong>C2</strong>: “I had a similar situation when a candidate rushed to the interview showing little interest. This made me think immediately that I would not wish to work with them. However, I had to force myself to keep calm and positive, to ensure the candidate is given sufficient attention.”</td>
</tr>
</tbody>
</table>
2.2 Understanding User Comments and Viewpoints

The collection of user comments is the starting point for the analysis. As mentioned before, the focus of this study is on emotions and body language. We refer to body language as the set of non-verbal behavioural cues. It instantiates instruments of communication, which a person adopts to express the emotions that affect him/her during the activity. These cues are transformed through the process of communication into social signals for other persons and conclude to the development (atomic or collective) of emotional intelligence[4].

Recognizing both emotion and body language cues comprise a set of two very important soft skills to be developed in interpersonal communication, particularly in a dyadic interaction such as in a job interview. Non-verbal communication carries most of the social meaning (about two thirds comparing with verbal communication). It illustrates emotional states, regulates the flow of interaction and provides valuable feedback to both actors in the activity. Awareness and recognition of behavioural cues accounts great value in social interactions[5].

Table 2 presents a set of comments provided by four different users watching the videos in our pilot system. The comments correspond to the beginning of the same job interview video, which includes actions such as entrance of the interviewee to the meeting room and handshaking. With regard to the focus of analysis we aim to capture key concepts on the emotion of the actors and the non-verbal behavioural cues. For example, in C3, some key concepts include: “handshaking”, “without manners”, “disrespectful”, in C4: “handshake”, “ignore”, “shake my hand”, in C5: “feel”, “discomfort”, “confusion”, “behaviour”, and in C6: “understanding”, “comfortable”.

<table>
<thead>
<tr>
<th>Comment text</th>
<th>Refers to</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>C3 : “Avoids the handshaking. Shows a person without manners, completely rude and disrespectful and maybe inappropriate for the job.”</td>
<td>Interviewee</td>
<td>u2</td>
</tr>
<tr>
<td>C4 : “I remember a situation when I offered a handshake and was ignored...I could not understand why. Later on, I realised what might be the reason. The person is a strict Muslim and me being a woman, it might not be permitted for him to shake my hand!”</td>
<td>Interviewee</td>
<td>u3</td>
</tr>
<tr>
<td>C5 : “The interviewer may feel discomfort and confusion due to the unexpected behaviour of the interviewee. The interviewer may be thinking that she would not wish to work with people who do not take her (or the job) seriously.”</td>
<td>Interviewer</td>
<td>u4</td>
</tr>
<tr>
<td>C6 : “She appears very understanding of the situation and tries to make the interviewee feel comfortable even though she is late.”</td>
<td>Interviewer</td>
<td>u5</td>
</tr>
</tbody>
</table>
Three types of diverse observations between the users are detected. The first type concerns the focus of observation with respect to the actors in the activity, i.e. the interviewee or the interviewer. For example, $u_2$ and $u_3$ focus on the interviewee, while $u_4$ and $u_5$ focus on the interviewer. The second type concerns the diverse approaches to characterize the same actor in the same part of the video. For example, although both $u_4$ and $u_5$ focus on the interviewer, $u_4$ points to a feeling of discomfort, while $u_5$ considers a sense of understanding of the situation. The third type concerns the context of comment, hence the interpretation of an event is affected by prior personal experience, e.g. the comment from $u_3$ corresponds to a personal real life experience.

Hereupon, we define an **individual viewpoint** as: *the focus and the collection of statements that a person develops when observing an activity*. In order to extrapolate the different viewpoints in the experimental study, each participant was given a questionnaire to complete prior to the interaction with the system. The questionnaire included quantitative as well as qualitative variables, which aim to extract measurements of their experience with job interviews and perceptions about the application of emotion and body language as communicative tools in the activity.

### 3 Semantic Approach for Comment Analysis

This section describes a semantic approach for analysing the user comments and extracting concepts, related to emotions and body language, with the correct meaning. The approach consists of three main steps, as shown in Figure 2.

**Text pre-processing** takes as input the comment in free text format and outputs linguistically tagged text. **Semantic pre-processing** inputs the linguistically tagged text and outputs a semantically tagged text with identified concepts. **Knowledge-statements extraction** uses the semantically tagged concepts and outputs knowledge statements extracted by queries over the ontologies.

The steps are explained below using C5 (from Table 2) as an illustrative example.
3.1 Text Pre-processing

The text pre-processing step comprises NLP techniques for text analysis using the Antelope NLP framework. Sentence splitting, sentence tokenization, Part of Speech (POS) tagging for each word and sentence chunking to extract meaning pieces of text inside the sentence that can stand alone, are performed using the Stanford parser for linguistic text analysis. This enables the linguistic tagging. Each word is assigned a POS tag and particular words are filtered out (e.g. articles and punctuation). Table 3 presents the processing components in text pre-processing step (left) and a short explanation of the resulted text structure that consists the output (right) and passes to the semantic pre-processing.

Table 3. The text pre-processing components and an illustration of the output

<table>
<thead>
<tr>
<th>Text pre-processing components</th>
<th>Example output [for C5 from Table 2]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokenization</td>
<td></td>
</tr>
<tr>
<td>POS Tagging</td>
<td></td>
</tr>
<tr>
<td>Sentence Chunking</td>
<td></td>
</tr>
</tbody>
</table>

3.2 Semantic Pre-processing

The semantic pre-processing includes the ontology-based word sense disambiguation (WSD) and linguistic semantic text expansion components. Two filters are applied for WSD, since each word along with the corresponding POS tag can have more than 1 senses in English language:

1. Selection of words and word senses according to specific lexical categories defined within WordNet to directly exclude those which are not significant for the application domain, i.e. job interview activity. Example lexical categories that are used for sense disambiguation as relevant to the domain are presented in Appendix.

2. Exploitation of the Suggested Upper Merged Ontology (SUMO), which provides direct mappings of English word units to concepts. From SUMO, 231 concepts out of 4,558 were selected as significant to the application domain, and the inclusion set has been validated with domain experts. The resulted concepts were used as word sense disambiguation indicators (second filter). Example SUMO concepts that are used as filters for sense disambiguation are presented in Appendix.

---

The linguistic and semantic expansion followed comprises of WordNet Lexicon queries, where synonyms and word lexical derivations were extracted to expand the word set, now in the context of the application domain. Furthermore, DISCO has been exploited to retrieve distributionally similar words from the Wikipedia corpus, and the semantic filters discussed above have been applied respectively. Table 4 presents the main processing components for semantic analysis and example output.

Table 4. The semantic pre-processing components and an illustration of the output.

<table>
<thead>
<tr>
<th>Semantic pre-processing components</th>
<th>Example output [for C5 from Table 2]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic processing Component</td>
<td></td>
</tr>
<tr>
<td>WordNet Lexical Category WSD</td>
<td></td>
</tr>
<tr>
<td>SUMO Concepts WSD and Identification</td>
<td></td>
</tr>
<tr>
<td>WordNet: Synonyms &amp; Derivations</td>
<td></td>
</tr>
<tr>
<td>DISCO Similar Words Extraction</td>
<td></td>
</tr>
<tr>
<td>The word due is removed from the text, as its sense has subsuming mapping to the SUMO concept Path: “a route along which motion occurs”. Contrarily, the word discomfort has two senses “the state of being tense...” and “an uncomfortable feeling...” which have subsuming mappings to the SUMO concepts: StateOfMind and EmotionalState respectively. From DISCO, the word appears to be strongly related with the word frustration that again has significant to the domain conceptual mappings as well as its derivation discomfiture. Similar results are returned from other words that correspond to valid domain concepts, e.g. behaviour is recognized as a TraitAttribute</td>
<td></td>
</tr>
</tbody>
</table>

3.3 Knowledge Statements Extraction

The final step consists of the knowledge statement extraction methods applied for the semantically filtered and linguistically enriched text structure. Regarding the focus of analysis of the activity (see Section 2.2), emotion and body language cues are in the focus of the methodology. Two ontologies have been developed. The WordNet Affect taxonomy of emotions has been translated to RDF/XML format consisting of 304 concepts related with subClassOf axioms. An ontology to conceptualize body language has been developed following a rich taxonomy of non-verbal behavioural cues and [4]. The ontology includes 15 concepts and 319 instances. Two properties are used to assert axioms: isExpressesBy relates <body language signal> [domain] (e.g. eye shrug, teeth grinding) with <body part> (e.g. eyes, arms etc.), <physical object> (e.g. pen, gum, tie etc.), <non-physical object> [range] (e.g. handshake); hasPossibleMeaning

relates <body language signal> [domain] with <body language signal meaning> [range] (e.g. frustration, defensiveness, interest etc.). Inverse properties have also been implemented. Figure 3 (a) and 3(b) show a small portion of the emotion and body language taxonomies respectively.

Each of the ontologies was pre-processed and index tables with concepts (and instances for the body language ontology) were constructed to increase querying efficiency. Reasoning was then performed for each identified concept in the text on the two ontologies to elicit potential knowledge statements. Table 5 presents the main processing components for knowledge statements extraction (left) and a short explanation of the output focused on the word “discomfort”.

Table 5. The knowledge statement extraction components and illustration of the output.

<table>
<thead>
<tr>
<th>Knowledge statements extraction</th>
<th>Example output [for C5 from Table 2]</th>
</tr>
</thead>
<tbody>
<tr>
<td>The word discomfort is used as example here to demonstrate the knowledge elicitation that its similar words mentioned above enable. Example ontology statements resulted from reasoning for the words discomfiture and frustration on the WNAffect and body language ontologies respectively include:</td>
<td></td>
</tr>
<tr>
<td>WNAffect – discomfiture reasoning: discomfiture “has equivalent” anxiety; anxiety “is a kind of” negative-motion; anxiety “is a kind of” Emotion.</td>
<td></td>
</tr>
<tr>
<td>Body Language – frustration reasoning: frustration “appears possibly when” nail_biting; “frustration” “appears possibly when” eye_shrug.</td>
<td></td>
</tr>
</tbody>
</table>
For each comment, the output of the analysis is a tree structure enriched with knowledge statements, as well as meta-data linking to modelling elements including: the video resource, the video snippet that the comment was added for, the user that added the comment, and indicators for the actor that the comment refers to and whether it corresponds directly to the video or to personal experience (Section 2.1).

4 Summary of the Output

To date, a total of 5 example job interview videos have been annotated by 10 users, providing in total 139 video snippets (for 8 job interview examples, as each video can have more than one interview example) and 193 textual comments. The set of textual comments was treated as a unified corpus for further analysis, as described in Section 3.

From the total of 193 comments, 127 were referring to the interviewee and 66 to the interviewer. 143 comments were related to the activity presented in the video (97 to the interviewee and 46 to the interviewer) and 50 to users’ personal experiences (30 as interviewee and 20 as interviewer). 152 comments were linked to emotion concepts, 168 to body language concepts, 144 to both and 17 to none.

From the analysis of comments, 174 unique words were extracted and linked to 274 unique concepts related to emotion (distinct 92) and body language (distinct 91). Each word was linked to concepts following the approach from Section 3

- as direct word (7.2% linked to emotions and 12.2% linked to body language);
- as a result of DISCO similarity (56% - emotions and 42.6% - body language);
- as a synonym (14.4% - emotions and 12.1% - body language);
- as a linguistic derivation (21.6% - emotions and 26.4% - body language).

As discussed in Section 3, each concept is linked with one or more SUMO domains according to its sense. Table 6 presents the five most frequent SUMO domains identified with example concepts from the corpus for both emotion and body language (refer to the Appendix for examples of the SUMO domain).

<table>
<thead>
<tr>
<th>SUMO domain (frequency (%))</th>
<th>Example concepts from corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Emotion</strong></td>
<td></td>
</tr>
<tr>
<td>SubjectiveAssessmentAttribute (25%)</td>
<td>despair, shame, encouragement</td>
</tr>
<tr>
<td>EmotionalState (23%)</td>
<td>anxiousness, confidence, euphoria</td>
</tr>
<tr>
<td>PsychologicalAttribute (15%)</td>
<td>wonder, humility, calmness</td>
</tr>
<tr>
<td>TraitAttribute (5%)</td>
<td>contempt, optimism, hostility</td>
</tr>
<tr>
<td>NormativeAttribute (3%)</td>
<td>oppression, approval, forgiveness</td>
</tr>
<tr>
<td><strong>Body Language</strong></td>
<td></td>
</tr>
<tr>
<td>SubjectiveAssessmentAttribute (24%)</td>
<td>pressure, negativity, upset</td>
</tr>
<tr>
<td>EmotionalState (7%)</td>
<td>nervousness, excitement, dissatisfaction</td>
</tr>
<tr>
<td>PsychologicalAttribute (7%)</td>
<td>combative, readiness, doubt</td>
</tr>
<tr>
<td>Artifact (6%)</td>
<td>body, pen, tie,</td>
</tr>
<tr>
<td>SocialInteraction (3%)</td>
<td>confidence, greeting, lying</td>
</tr>
</tbody>
</table>
Based on the questionnaire, the experience of each participant with the interview process was indicated (participants were asked of the number of interviews undertaken both as an interviewee and as an interviewer using a categorical scale with values). Users were also asked to indicate the importance of emotion and body language in the job interview activity. Out of 10 users in the study, 1 had much experience as interviewee (over 15 interviews) and 3 had much experience as interviewers (over 15 interviews), 9 were not much experienced as interviewees and 7 not experienced as interviewers. 7 of them replied that emotion is important, while the users that gave negative answer or they did not know, were 1 much experienced as interviewee and nowise as interviewer, and 2 with little experience as interviewees but much experienced as interviewers. 9 of them replied that body language is important, while the one that did not know was much experienced as interviewer.

90% of the comments related to personal experience and 65% were referring to the interviewee were contributed mostly by users with either little experience as interviewees or interviewers, and similarly, the same class of users provided the highest proportion of video related comments (70%) and comments referring to the interviewer (78%). These users also contributed the highest rate of comments linked to emotion (73%) and body language (72%).

Comments from users with little experience as interviewees were mainly linked to the SubjectiveAssessmentAttribute (24%), EmotionalState (14%), and PsychologicalAttribute (10%) domains from SUMO. Comments from those with much experience as interviewees were linked with the same domains but in highest rates (approximately 2-3% additional rate). Users with much experience as interviewers, contributed also comments mostly to linked to these domains in highest rates (over 3% additional rate). Other domains were also included in the resulted sets for experienced interviewers, e.g. SocialRole (2.5%), BodyMotion (1.5%) and SocialInteraction (3%), whilst not so important—in terms of frequency rates—for the classes of users mentioned above. Overall, users with experience as interviewers commented about the interviewers, while the other users tend to focus on interviewees. This indicates that people recognise body language and emotions related to the role they have experienced most and miss these aspects in the role they have not had experience in. We are currently further processing the data, together with collecting more comments and user profiles. If this hypothesis is confirmed, this will give an indication that users’ experience with body language and emotions may be related to the comments they make.

5 Discussion and Future Work

Much work has been done in technology-enhanced learning, focusing on intelligent environments for experiential workplace learning. Job-related experiences are captured through these environments for organizational knowledge [7] or every-day computer based tasks in work promoting self regulated learning [8], and in academia writing skills are being developed by sharing students’ experiences [9]. Records of job-related activities (e.g. videos) have been used in [10] to create pedagogical
scenarios for experiential learning. We aim to distinguish from these projects in four points: include multi-viewpoints in the activity model; advance the knowledge elicitation process by implementing methods to provide user-awareness of related activities; provide more expressive models to augment digital content; and test augmented video resources in simulated settings for learning. Furthermore, this work contributes to a new stream in user modelling utilizing ‘real-world’ work context models to improve adaptation[11], and using digital traces from social content to derive user profiles [12]. Instead of explicit user profiling, we will provide a mechanism for deriving an extended context model which preserves different viewpoints on an activity, and can be used to improve adaptation, as well as a source for clustering and profiling users. Similarly to [13], we focus on awareness and recognition of social signals to empower adaptation, but we are applying it to job interviews where diverse interpretations should be catered for.

Emotions and body language have a strong presence in the corpus collected, following the indications given to users prior to system interaction. The results gathered so far empower the feasibility for context capturing. The expressivity of body language ontology will be increased by redesigning concepts, instances and relations and the information extraction algorithm will be refined accordingly, regarding also the linguistic ontology pre-processing as needed. So far the SUMO concepts have been validated only for precision, while recall has to be addressed. The plan is to explore further the relations in SUMO with rule based reasoning techniques, identify recalled concepts and validate the consistency of the sub-ontology. Similarly, the next step is to formally evaluate the NLP tools used and the corresponding implemented algorithms.

A more uniformed corpus has to be collected in terms of user profiles to normalize the distribution of experiences and conclude more explanatory results. The extraction of individual viewpoints appears feasible, however, more thorough analysis has to be undertaken in order to identify possible relations, e.g. between the experience of participants and the concepts identified in the corpus, derive patterns between the concepts extracted from comments and shape the activity model according to time and context. Evaluation steps will include validation with domain experts and exemplification of simulated context queries for content retrieval, involving users from the research study.

**Acknowledgement**

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no ICT 257831 (ImREAL project).

**References**

[1] A. Mitrovic and A. Weerasinghe, “Revisiting Ill-Definedness and the Consequences for ITSs,” presented at the Proceeding of the 2009 conference on Artificial Intelligence in...
Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling, 2009.


Appendix

Examples of WordNet lexical categories and SUMO concepts used for word sense disambiguation and filtering.

<table>
<thead>
<tr>
<th>WordNet lexical categories</th>
<th>SUMO concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>[noun.act]: acts or actions</td>
<td>[SubjectiveAssessmentAttribute]: a kind of normative attribute for a subject</td>
</tr>
<tr>
<td>[noun.artifact]: man-made objects</td>
<td>[SocialInteraction]: interactions between cognitive agents such as humans</td>
</tr>
<tr>
<td>[noun.attribute]: attributes of people and objects</td>
<td>[SocialRole]: specifies the position or status of a cognitive agent (as human) within an organization or other group</td>
</tr>
<tr>
<td>[noun.body]: body parts</td>
<td>[Agent]: something or someone that can act on its own and produce changes in the world</td>
</tr>
<tr>
<td>[noun.cognition]: cognitive processes and contents</td>
<td>[SubjectiveAssessmentAttribute]: attributes that are specific to morality, legality, aesthetics, etiquette, etc. Many of these attributes express a judgment that something ought or ought not to be the case</td>
</tr>
<tr>
<td>[noun.communication]: communicative processes and contents</td>
<td>[Artifact]: an object that is the product of a making</td>
</tr>
<tr>
<td>[noun.feeling]: feelings and emotions</td>
<td>[BodyMotion]: any motion where the agent is an organism and the patient is a body part</td>
</tr>
<tr>
<td>[noun.motive]: goals</td>
<td>[EmotionalState]: the class of attributes that denote emotional states of organisms</td>
</tr>
<tr>
<td>[noun.person]: people</td>
<td>[StateOfMind]: transient features of a creature's behavioral/psychological make-up</td>
</tr>
<tr>
<td>[verb.motion]: walking, flying, swimming</td>
<td>[Perception]: sensing some aspect of the material world</td>
</tr>
<tr>
<td>[verb.perception]: seeing, hearing, feeling</td>
<td>[FormalMeeting]: any meeting which is the result of planning and whose purpose is not socializing</td>
</tr>
<tr>
<td>[verb.cognition]: thinking, judging, analyzing, doubting</td>
<td>[BodyPart]: ...small components of complex organs</td>
</tr>
<tr>
<td>[verb.communication]: telling, asking, ordering, singing</td>
<td>[PsychologicalAttribute]: attributes that characterize the mental or behavioral life of an organism</td>
</tr>
<tr>
<td>[verb.contact]: touching, hitting, tying, digging</td>
<td>[TraitAttribute]: attributes that indicate the behavior/personality traits of an organism</td>
</tr>
<tr>
<td>[verb.emotion]: feeling</td>
<td>[RegulatoryProcess]: an guiding whose aim is the enforcement of rules or regulations</td>
</tr>
</tbody>
</table>
Recommender Systems and the Social Web

Amit Tiroshi, Tsvi Kuflik, Judy Kay and Bob Kummerfeld
Abstract. In the past, classic recommender systems relied solely on the user models they were able to construct by themselves and suffered from the “cold start” problem. Recent decade advances, among them internet connectivity and data sharing, now enable them to bootstrap their user models from external sources such as user modeling servers or other recommender systems. However, this approach has only been demonstrated by research prototypes. Recent developments have brought a new source for bootstrapping recommender systems: social web services. The variety of social web services, each with its unique user model characteristics, could aid bootstrapping recommender systems in different ways. In this paper we propose a mapping of how each of the classical user modeling approaches can benefit from nowadays active services’ user models, and also supply an example of a possible application.

1 Introduction

Information overload is a phenomenon that has invaded every field in our lives, from work activities (decide which books to order, which emails to read first) to leisure time ones (which movies to see, which restaurants to go). One way to ease the problem is through the use of recommender systems [1], systems that try to match users and items/entities that might interest them. There are several classic approaches for generating recommendations: collaborative filtering [2], content-based [3], case-based [4] and hybrid methods [5]. Most recommender systems require a user model to base their recommendations on and every method described earlier requires a different type of user model. Until a decade ago each system had its proprietary user model, however with the bloom of the internet and connectivity, user models sharing and bootstrapping from online sources are becoming a real possibility. One possible source for bootstrapping user models is the freely available personal information from the social web. Social web services are online services that let their users connect, communicate, share and collaborate with others. Users can link themselves to groups, individuals and causes, they can share all types of content (written, visual, audio) and they communicate both live and in a delayed manner. Each social web service has its unique characteristics which are also reflected in its user model, some let users define
their interests explicitly as a set of features (Facebook1, Linkedin2) others do so implicitly and in plain text (Twitter3, Blogs). Facebook only allows a bidirectional connection among users (if user A is connected to B then B is also connected to A) while Twitter users can follow without being followed (user A is linked to B, B is not linked to A). As a result, the social web contains vast amounts of personal information about users that is free and publicly available or can be made available by the users. This information may serve as a source for information used by online recommender systems to bootstrap their user models and to solve the “cold start” problem. In this paper we survey existing social web services and show how the different recommendation approaches (or user model representations) can each benefit from the social web’s available user models, and present an example in the form of a possible application.

2 Background and related work

The recommendation approaches mentioned earlier are the classical ones for handling information overload [6]. Each of them has a unique method for modeling its users’ interests and how to match items accordingly. In this section we will review the various models, starting with the collaborative filtering approach (CF). CF [2] is based on similarity of user preferences, it assumes that users that agreed in the past on items they liked will probably agree on more items in the future. For example, taking one user’s bookshelf and crosschecking it with shelves of other users, finding those with similar books will yield several possible book recommendations for that user. To carry out such an operation, ratings of items must be gathered and stored from a large number of users. This approach is called user-user CF. A variation of CF is to base filtering on similarity of items (item-item CF) rather than similarity of users. A matrix exists, which represents the relationship between each pair of items. Thus, every item listed under the active user can serve as a lead for potential related items in the matrix. Overall, the general user model of CF systems requires a matrix of users’ ratings of items. Having such a matrix is not easy and a major challenge is how to support a new user (new user problem) or how to rate a new item (first rater problem) – both are two aspects of the “cold start” problem of CF.

In the content-based approach, recommendations are made based on content analysis. The content is a set of terms representing an item (Website, Document, Email) or describing it (Movie/Music CD/Restaurant descriptions), usually extracted from the larger textual description of the item. To create a user model, the content that interests the target user is either explicitly given or implicitly learned through machine learning techniques [7]. Then, when new content becomes available, it is analyzed and compared to the user model, and recommended (or not) to the user based on that similarity. Among the most common techniques used for content similarity analysis is the vector space model that uses TF*IDF [8] weighting, Rocchio

1 http://www.facebook.com
2 http://www.linkedin.com
3 http://www.twitter.com
algorithm [9] and the naïve Bayesian classifier [10]. Another approach, quite similar to the content-based approach is a feature based one, where users (and items) are represented by preferences of specific features (like movie genre, book author etc). Again, these features form an n-dimensional vector where similarity of users and items may be measured by a cosine in an n-dimensional space.

The third approach, case-based [4], is another variation of the content-based approach, aimed at generating better recommendations for feature described items such as consumer products, based on past interactions with the system by similar users. In this approach, user sessions are recorded and when similar users request recommendations – similar sessions generated by similar users (users with similar preferences), are used as a basis for recommendations. An exemplifying implementation is presented in [11].

Hybrid recommender systems [5] are systems that combine two (or more) approaches together in order for them to overcome each other's shortcomings. For example, a system that combines the collaborative filtering approach with a content-based recommender can overcome the first rater problem by matching new items using content analysis, as demonstrated at [12].

The Social Web was introduced in [13] as a project in which people could create an online representation for themselves, get organized in groups and communities, share knowledge and items while interacting and collaborating with others. Since then services implementing those concepts have evolved and currently many variations can be found. Among the commonly adopted ones are Facebook, Twitter, Flickr⁴ and Blogger⁵. Facebook is a social web service (also categorized as a Social Networking Service – SNS) that focuses on personal life aspects, its users are able to create an online rich representation of themselves, containing elements varying from detailed profile attributes to personal photo albums and status sharing. Facebook users are encouraged to connect and interact with others whom they know and join groups based on shared interests. Since then services implementing those concepts have evolved and currently many variations can be found. Among the commonly adopted ones are Facebook, Twitter, Flickr⁴ and Blogger⁵. Facebook is a social web service (also categorized as a Social Networking Service – SNS) that focuses on personal life aspects, its users are able to create an online rich representation of themselves, containing elements varying from detailed profile attributes to personal photo albums and status sharing. Facebook users are encouraged to connect and interact with others whom they know and join groups based on shared interests. Flickr on the other hand is a social web service designed for photography hobbyists and professionals, its users can upload their works and share them publicly or to specific interest groups. Other users can then comment and tag elements in those photos and socially interact. Blogger is a service that allows its users to log their thoughts and happenings online and share them with others. Such "Posts" can then be commented on by other users, leading to social interaction. Twitter is also a blogging service, however for micro posts which do not exceed 140 characters. Twitter is characterized by an additional informal usage pattern in the form of frequent real-time updating. The user models of these services are accessible to 3rd parties through APIs with the users' consent. On the next section we will map between those services and their possible contribution to classical recommendation approaches user models. The services mentioned above and in the next section are the leading representatives of current social web services (usage wise). The given map can be used to project additional existing services' contribution to user models based

⁴ http://www.flickr.com/
⁵ http://blogger.com/
on their similarity to the chosen services. For example LinkedIn\(^6\) is a social web service that shares a similar concept to Facebook but aimed at connecting professionals instead of social friends. Thus in the approaches Facebook is mapped to, it would also be suitable to use LinkedIn if the context of the application is more "Professional" then social. Additional relatively less adopted social web services are: Plurk\(^7\) and Tumblr\(^8\) which belong to the micro-blogging category Twitter is part of. There are also many photo sharing services\(^9\) and blogging services\(^10\). One service that stands out is YouTube\(^11\) being an exact match to Flickr in the map presented on the next section both in scale and properties with the only difference being that it serves videos instead of photos, a difference that does not affect its contribution relatively to Flickr.

Social web services have been used for bootstrapping user models. In an early study \([14]\), social web service profiles were captured and mapped to a “Taste Fabric” using ontologies of books, music, movies and more. The taste fabric was constructed using machine learning techniques to infer semantic relevance among the ontologies. It was then used to recommend new items to users who share the same cluster of “Taste”. In another study \([15]\) bootstrapping a Scrutable User Modeling Infrastructure (SUMI) from fragments of the user’s user model located at various social e-networking and e-commerce domains was explored. Using APIs from Facebook, Amazon, eBay and Google OpenSocial, SUMI was able to harness users’ data for its own purpose of Lifelong User Modeling and personalized learning.

3 Mapping Social Web Services Contribution to Classical Recommender Systems Approaches

Social Web Services, by nature, contain large amounts of personal information about their users. Some details may be publicly available while more may be kept private and released explicitly by the users. This information can be valuable for online recommender systems seeking to bootstrap a user model for first time users, in order to overcome the “cold start” problem, where without any personal information (or interaction history) the system is unable to provide a personalized service to the new user. It can also be used to enrich existing models with complementary data from different domains.

\(^6\) http://www.linkedin.com/
\(^7\) http://www.plurk.com/
\(^8\) http://www.tumblr.com/
\(^9\) http://en.wikipedia.org/wiki/List_of_photo_sharing_websites
\(^10\) http://en.wikipedia.org/wiki/Category:Blog_hosting_services
\(^11\) http://www.youtube.com/
Figure 1 illustrates the leading social web services and their possible contribution to the user models of classical recommender systems. We will now analyze the specific contribution to each approach starting with the CF approach. Since CF relies on user ratings of items, bootstrapping those ratings from the social web services could have a tremendous contribution. The networks offering information that resembles such ratings are Facebook, Twitter and Blogger. On Facebook, users can explicitly declare their interests through profile features, association with groups and fan pages or through status line updates. Such attributes once extracted can be mediated to ratings on items, for example: a user linking her profile to ‘Levis’ fan page is essentially rating the brand and its products as favorable. The same process can be used for tweets (the name for a Twitter post), however methods such as sentiment analysis are required in order to resolve the precise rating, since an open text sentence regarding ‘Levis’ for example, can be a statement of endorsement or of hate. Flickr being a visual content sharing hub is less helpful in the interests bootstrapping process and thus not linked in the mapping above. A second possible contribution social web services could have for CF is related to the social links they store as part of their user models. Social links might serve as an indicator of trust among users, and trust could be an important factor among raters in a collaborative filtering system. In a research by a collaborative filtering system is demonstrated, in which users can request recommendations based on items rated by specific users whose ratings they trust. Facebook’s social links along with the mutual interests of the two people connected could supply this trust factor. On Twitter the people a user follows can serve as raters in whom she trusts on the specific subjects tweeted about.

Content-based recommender systems require a set of terms representing the content the user is interested at. These terms can be extracted from the user’s social web service profile, in which the text tends to be short and focused. Such interest terms can be extracted from Facebook fan pages and groups the user is associated with, the group/page names themselves are suitable (as in the ‘Levis’ example) and additional terms can be found in the accompanying short descriptions. Additional
short open text fields that could ease the term identification process are status lines and wall messages in Facebook, Twitter's messages which are limited to 140 characters are appropriate too. Blogger posts and blogs in general are more extensive in content then the services mentioned above, therefore their contribution would be similar to classical content based sources. Their advantage is that they already contain the user’s content of interest organized in a single point of access, hence serve as a more comfortable bootstrapping source. Although Flickr is not a textual site, the tags used to annotate images (addressed in Figure 1 as "Content Classification") can possibly serve as focused terms of interests for the sharing user, this approach was explored by [19] which also surveys additional similar methods. Content classification (aka "Tagging") also exists in all the mentioned social web services and could be used in the same way mentioned, on each service the content elements which are "Taggable" varies.

Case-based recommender systems having their origins in content-based ones can benefit from social web services in the same ways mentioned above. A possible unique contribution of social web services to case-based recommendations could be in the form of bootstrapping feature weights. Instead of requiring users to rank features based on their importance (for example price vs. color), those can be retrieved using stereotypically matching user profiles to predefined weight vectors. For example if an online consumer recommending system has mapped their products to various consumer stereotypes (students vs. professionals) and set for each stereotype a preset of feature weights, now all that has to be done is find whether a user is a student or professional, a detail that is available on a social web service such as Facebook.

Hybrid recommender systems can benefit from the fact that some users have their social web service profiles linked together, hence having different representations of the user that can complete each other without needing to manually link between the two system profiles (identity linking). A user that is both a member in Facebook and Twitter for example, and has those two profiles connected using methods similar to such mentioned at [20], can permit a hybrid recommender system to use the first to bootstrap its CF user model part and the second to bootstrap its content-based user model. In case the user's social profiles were not priorly linked the recommendation system can attempt to link them automatically by using personal details features available on both social web service as shown in Figure 1, there are commercial social data aggregation services which do this, for example ZoomInfo\(^{12}\). Another option for hybrid systems to enrich their models is to use classified content with identical tags across services, for example photos of a user from Flickr can be matched to textual items from other services and users, that were tagged identically, thus aiding in bootstrapping a content based user model. An issue which requires attention when using social data from multiply sources is user modeling interoperability. Each source’s user model can have its unique data representation and formats, leading to a need in translation/conflict resolution/mediation methods that could integrate them all into a unified model. Such methods were surveyed in depth in a recent study [21].

\(^{12}\)http://www.zoominfo.com/
4 Theoretical Use Case

To illustrate the potential benefits which were described in the previous section we would like to propose a theoretical example of a socially enhanced museum guidance system. The purpose of the system would be to offer personalized museum tours tailored to users' interests as reflected by their social data. Systems for personalizing the museum experience were studied in [22] [23] [24] [25] and more, various mechanisms were required to initialize those systems' user models. In the proposed approach all that is required is visitors consent for the museum's personalization system to access their online social web profiles in order to bootstrap a local user model. Once the museum's system has access to the various social profiles of the visitor and its local user model is bootstrapped, exhibits of interests can be recommended using any of the classical approaches. Actual links that were manually found between real life exhibits presented in the Hecht Museum\(^\text{13}\) and public social profiles are also attached to demonstrate the approaches suggested.

A content-based exhibits recommendation method for example would use the user's Twitter stream as a source for terms of interests. The terms would be extracted using a method such as Bag of Words [8], and then matched against content describing the museum exhibits using content analysis methods. If the user had tweeted about cosmetics and the museum hosts exhibits related to that they would be recommended for a visit (Figure 2 and 3).

![Fig. 2](Image)

**Fig. 2.** A Twitter post (Right) about cosmetics and a related exhibit (Left) in Hecht Museum that could be recommended to its owner

A different approach that can be combined with the one mentioned above would make use of the users' social profiles and log of visited exhibits to personalize future visitors experience based on CF. If a user visited certain exhibits and her/his Facebook page mentions she/he is a "Fan" of certain items, those would be saved for later matching

\(^{13}\text{http://mushecht.haifa.ac.il/}\)
against new visitors profiles. New visitors would be recommended exhibits that were viewed by people whom they most resemble based on the items they are “Fan” of. Another interesting case for social web services user model bootstrapping would be in hybrid recommender systems. An exemplifying scenario would be museum visitors that have taken photos of exhibits they have seen and tagged them, those can then serve as a basis to identify visitors with similar interests (using the CF approach on social profiles) and recommend those tagged exhibits or similar ones, based on content-base recommendation. The recommendation process would be in the lines of: Find user profiles resembling current visitor's profile, extract tagged photos that are also related to museum's key terms, recommend exhibits relating to those. The great advantage in this case is in the fact that the two user models (CF and content based) are already linked together through the social web services, thus identity linkage is not required. Also such links between profiles allow users to maintain their partial models in the services fitting best for them, for example a user using photos could store them in a service that specializes in it such as Flickr, and link that profile to a Facebook profile which is more suitable for maintaining social relations online.

Fig. 3. A visitor whose Facebook profile (Left) states he works in a Maritime Archaeology Unit might be recommended the Ma’agan Mikhael shipwreck exhibition (Right) in Hecht Museum

Finally social web service based recommender systems can also contribute to future uses whether they will be using the same system or 3rd party systems by asking permission from the users to update their online social profile with information related to their latest usage. In our example this would be done by asking visitors for permission to update their Facebook/Twitter streams with summaries of the tour they have taken, e.g. a list of exhibits visited and personal photos taken with them, relevantly tagged. It could enrich the users’ experience by giving them a memoir of their visit, and also serve other museum systems in knowing which exhibits to recommend to them.

5 Discussion

This paper surveys social web services and presents a mapping between them and possible usage of their data to enhance classical recommendation approaches’ user models. We have also presented a theoretical example for a recommender system that is based on the mentioned methods, and illustrated on real publicly available social
data how it can be linked to actual exhibits. Future work will focus on concrete evaluations of the methods proposed. Also we would like to extend the mapping to modern recommendation approaches such as Social Tagging Based and Group Based.

Social data usage comes along with the responsibility to preserve its owner’s privacy. Besides the elementary rules of using the user’s data only for the purposes to which permission was granted and not to forward it to unauthorized parties, there are also some less obvious rules that should be taken into account (e.g., for how long can the data retrieved from a social web service be stored by a recommendation service? This is important in order to prevent the service from using outdated data that could lead to misleading/offensive recommendations). It was out of this study’s scope to cover this issue, however a future study should offer a mapping of privacy risks and preservation techniques corresponding to the utilization approaches suggested.

6 Bibliography

13. Hoschka, P.: CSCW research at GMD-FIT: from basic groupware to the social
Towards a Digital Learner Identity

Adriana Berlanga and Peter Sloep
Towards a Digital Learner Identity

Adriana J. Berlanga, Peter B. Sloep
Centre for Learning Sciences and Technologies (CELSTEC)
Open Universiteit Nederland
{adriana.berlanga, peter.sloep}@ou.nl

Abstract. This paper presents a first attempt to describe the relevance and importance of a Digital Learner Identity in the context of supporting lifelong learners. A Digital Learner Identity is an augmented user model that captures the traces lifelong learners leave in their digital, social and physical worlds, all combined in a single model. In this paper we start looking at the notion of offline identities, and how it is changing because of the emergence of the social web. Then we focus the lens of our attention on how a unique Digital Learner Identity could help learners to find people to collaborate with (even people they are not linked to), and to understand how social networks (digital or physical) impact the way they learn. Finally, we describe a scenario to exemplify how the proposed Digital Learner Identity could support lifelong learners.

Keywords: Digital identity, digital learner identity, lifelong learning, professional development, augmenting user models, learning analytics

1 Introduction

Consider the following scenario:

Judith works as translator in a multinational company. She translates manuals for medical equipment into different languages. Often she seeks recourse to the Web to look for technical terminology, to participate in communities of translators and ask them for help, but also to support others. She is an avid user of social media applications for her job but also privately, especially regarding her passion, photography. She uses Flickr, for instance, to show-case her pictures. Through it, she receives reactions from many people, gets regularly invited for the ‘picture of the month’, and participates in communities of photographers. She also regularly comments on pictures by others and participates in various other kinds of online interactions. Judith takes photography courses at a local art centre and is member of an association of photographers that organises visits and activities.

As it happens, Judith feels ever less motivated by her job as a technical translator. She would like to do something else for a living, something she genuinely enjoys and can be more passionate about. So, she would really like to make a living as a photographer. Unfortunately, she has no clue where to start: what kind of jobs could she find that suit her, what kinds of competences do these require, and whom can she contact to find this out?

Judith is the exemplary lifelong learner, someone who finished her formal, compulsory education some time ago and needs ways to keep up with her job or, indeed, wants to chart out other job avenues [1]. The traditional solution to her
predicament would be to assume the existence of a competence map and some sort of almost mechanistic evaluation process that allows her to (i) compare her desired competences with her current ones, and (ii) receive a recommendation for what learning activities and resources she should consider to fill her competence gaps. However, because this is a domain or task-centred approach only, it does not work for her. Her profile is a rich one and does not only include her formal qualification as a translator - for which the traditional approach could work - but also her informal ones as a photographer. So the traditional approach does not do justice to her full profile qua lifelong learner.

To avoid the pitfalls of the traditional approach, we highlight in this paper that lifelong learners leave a great number of digital traces about their learning behaviour. We argue that these should all be considered in a single, unique identity, which we have called a Digital Learner Identity. With it, lifelong learners could find out how to acquire support on finding out what their talents are, on what they have learnt, on what they maybe should reflect on, and on how they could share knowledge and generate creative and innovative ideas. The backbone idea, which we will further develop in the paper, is that many if not most lifelong learners unknowingly already have a Digital Learner Identity and that over time this will become increasingly richer. In our approach we go further than current work in open learner models [2]. Crucially, we intend to capture the learner identity of the lifelong learner without targeting a predefined domain or task-related requirements.

This paper is structured as follows: Section 2 discusses why the notion of identity is changing. Section 3 explains the benefits of collecting profiling information for constructing a Digital Learner Identity, and Section 4 presents future work.

2 The Notion of Identity

The notion of identity has always been inextricably linked to the notion of a spatio-temporal individual, that is, an individual that is bounded in space-time. Our identity is a complex characteristic of us qua spatio-temporal individuals, which comprises our beliefs, desires and dispositions. Psychologists are mainly interested in our self-identity: our beliefs and desires, and in how they cause us to act in particular ways. Sociologists are interested in our social-identity: how our acts affect others, who are likely to have different beliefs and desires. These two are of course intimately connected, but nevertheless distinguishable.

Swann [3] has introduced the term identity negotiation to help us understand how our identities come about and change in social interactions. Others have expectations of us and we ourselves have an image of who we are and what we are capable of (our self-identity). In our contacts with others, these two images of our identity are confronted with each other. The result of the negotiations should be that expectations and self-identity start nearing each other to become congruent. If that happens, it may form the starting point of productive collaboration as the other roughly knows what she can reasonably expect and we do not feel under or overtaxed.

Self-identities develop over time. They develop both through social negotiation with others and in inner dialogue with ourselves. People expect others to maintain a relatively stable identity and, indeed, there is a tendency to maintain a stable self-
identity across time and across different groups. To the extent we succeed in doing so, we may spend time productively rather than on investing in developing and maintaining different identities.

2.1 Offline and Online Identities

With the proliferation of social media most people 'go online' and thus maintain an online identity, however elementary. On a hotel reservation site commenting on the quality of a particular hotel that you stayed in already constitutes the barebones of an online identity. Twittering regularly or writing a blog make for much richer online identities. Interestingly, since comments go to, for instance, hotel.com and a blog is maintained in blog.eu, people by default acquire multiple online identities. One would have to use the same unique username across sites to bring some unity to one's online identity. Nevertheless, the process of negotiating one's social identity is different online than offline. The concept of negotiation constraints is, therefore, useful here.

First, negotiating one's offline social identity is constrained by being one unitary spatio-temporal individual. Since negotiations are done face to face, it is physically impossible to have two identities at the same time. Online, though, one can have many identities simply by using different usernames. This is common practice and really only a problem in cases in which such 'double' identities are somehow misused.

Second, negotiating one's identity is also constrained by moral rules. As it is much easier to adopt two identities online than offline, moral constraints are also more easily evaded. So the negotiation of online social identities is significantly different than the negotiation of offline social identities. These two identities need also to be congruent with one's self-identity, in order that it frees up resources for productive work. Online negotiation costs are obviously smaller than offline negotiations. Thus online friendship comes ‘cheaper’ than offline friendship, as it takes one mouse click to befriend someone [4].

2.2 Online Learner identities

What does all this imply for somebody's online identity as a learner? In the network and knowledge society, learning has taken an altogether different shape. Lifelong learning now is the adage. People should keep learning, and compulsory education should prepare them for that. Lifelong learning has become joint knowledge sharing and knowledge creation rather than top-down knowledge transfer [1]. The social web offers unprecedented opportunities for this kind of learning [5]. In this context digital learning identities become of paramount importance. They should be the central hub of any lifelong learner's online activities. It is that identity also that is part of the continuous identity negotiation processes, with learning, i.e. knowledge sharing and development. This prompts us to make two observations.

First, it is clear that learners want to enter their identity negotiations with their fellow learners with only one such identity. For lifelong learners, their professional identity and learning identity are wedded together, the one feeding into the other and vice versa. Having a fragmented online learner identity is detrimental to the aim of sharing knowledge and having creative interactions with the most suitable people around. Second, this online profile should coincide with one's offline profile. As we
already indicated, a stable identity is conducive to being productively engaged with others.

Unfortunately, the way the social web currently operates promotes fragmentation of online identities. It is not in the interest of commercial social media sites to promote the use of consistent identities across sites. Indeed, they act as walled identity gardens. And to the extent that online learning makes use of these sites for profile information, digital learner’s identities therefore are fragmented. What can be done about this? The only answer seems to be to provide an alternative profiling service, one that upholds a Digital Learner Identity.

3 The Benefits of a Digital Learner Identity

Siemens [6] defines Learning Analytics as the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs. For educators, this kind of information has implications for how one perceives teaching, learning, and assessment [7]. For lifelong learners it has implications not simply for monitoring one's own learner performance, but also for how one perceives the learning process. This implies that, tacitly, learning analytical data are collected from learners’ online and, if possible, offline actions, from formal and informal educational settings, as well as from their behaviour in online and offline social interactions (e.g., blogs, social media participation, face-to-face interviews). This requires combining data from different sources, in different formats, collected using different techniques.

A Digital Learner Identity as a model not only should include this information, it must also have mechanisms to feed the data back to the learner and allow her to modify them. Collecting this social information has benefits for lifelong learners; because of space constraints, we describe here only two of the most relevant ones.

First, a Digital Learner Identity would allow learners or software agents to identify relevant (groups of) people with whom they could make contact. It could do so in two ways: (a) selecting people or groups thereof by link analysis or interconnectedness with others, or (b) finding groups that are very similar to one another, but may never actually interact online. The later are referred to as abstract groups [8]: groups in which the members do not interact explicitly, but the online and offline traces of which demonstrate cohesiveness in some way. Identifying these abstract groups can help lifelong learners find people whom they can ask for help or advice [9].

Second, a Digital Learner Identity is useful to find out how social networks influence learning. Christakis and Fowler [10] state that to know who we are, we must know how we are connected with others in our social network. The influence of social networks can work in two ways: the structure of the network (connections), and the information, behaviour that it is disseminated throughout the network (social contagion). Research on the impact that social networks have on everyone's lives shows that behaviour and ideas may spread because of social network connections alone, no external driving forces or an internal zeal to imitate someone are needed. Social networks influence, for instance, happiness, loneliness, cooperation, and even obesity. This influence spreads through the social network as far as to third-degree connections [10].
These observations have a tremendous impact on the way online learning should be organised. For one, it is safe to say that learning is an attitude that is contagious throughout one’s social network; note how it is the Digital Learner Identity that drives this contagious attitude towards learning. But second and at a more profound level, the way a network for learners [11] is structured matters. Importantly, small-world communities foster collaboration; a balance of power in the form of the lack of few yet powerful, i.e. well-connected individuals fosters self-organisation [10]. These network characteristics are conducive to social learning, knowledge sharing and knowledge creation. And precisely because of this, in earlier work we introduced ad-hoc transient groups (or communities) as means to seed the emergence of networked communities and forge network structures that are conducive to learning [9].

4 Future Work

Returning to our scenario, we can now easily detail how the availability of a Digital Learner Identity could benefit Judith:

In her photography course Judith gets an assignment about Maastricht, the city in which she lives. She needs to create a conceptual idea of the city, marking touristy spots for a travel agency. She first looks on the Web to acquire her first ideas of what she could do. She subscribes to some interesting pages, follows some Twits, and saves some links in her Delicious account. Then she goes out and takes many pictures of her city, sits in a cafe, and opens her digital learner identity application. The application asks her which information should be kept. She selects the pictures she likes the most and decides to keep all her traces she left on the Web, some of the routes she followed through the city, and indicates which information is public. Then the application tells Judith that two of the pictures she took were also taken by other people. One of them, Ana, lives in Madrid and leads a marketing bureau. Judith then clicks on Ana’s Learner Identity and sees the relations they have in common, this includes social relations but also affinities they share, or activities they both have engaged in. The profile also includes the competences Ana has, her experience, formal education, job history, and her current position of marketing director. Judith clicks there, and gets a graph of people that work in this type of business -who are in her social network and whom she could contact – alongside with topics related to this type of work, recommendations of which (online) courses fit her best to get acquainted with this job, online and offline communities that could be of her interest, and places she could visit to learn more. These recommendations already take into account that Judith speaks Spanish fluently and has technical translator skills. At this point she asks herself ‘What if I were to open a marketing bureau myself?’

Simple as it may seem, a lot of work still needs to be done to make the scenario just described a reality. First, the Digital Learner Identity should be automatically updated with dynamically augmented data, with the ‘tracks and traces’ learners leave in a variety of social media sites and in real, physical situations. Only then can one be sure this information adds to a rich, varied and ever up-to-date digital learner identity.

The next, possibly even harder challenge is how these data once stored in the learner’s Digital Learner Identity could be used, exploited and visualized to understand how the individual is learning while connecting and interacting with other
people. At the simplest level, they will provide a means to recommend relevant resources (digital resources, people). At a more interesting but also more complex level, they could be used to infer learning patterns or models, and automatically derive descriptions of competences learners have acquired while interacting in social learning contexts.

On the roadmap for the further development of the Digital Learner Identity idea the need to explore existing technologies features large. Such technologies can be found in the areas of generic user modelling [12]; generic user model ontologies such as GUMO [13] – in order to interpret distributed user models, combining static, dynamic and ubiquitous information; aggregation of profile information from different online services such as Mypes [14]; and authentication and authorization protocols so learners can control and manage their identity profiles, such as OpenID2.0 [15]. This exploration should lead to an understanding of what could be relevant for the Digital Learner Identity, and, importantly, to find out what still needs to be developed. Besides the technological solution, we will elaborate further the social learning meta-model and how it should be defined and built to support the Digital Learner Identity.

In the long run we want to investigate how the Digital Learner Identity could capture emotions and affective states, as well as explore business models that assure the sustainability of the Digital Learner Identity. The latter requires one to consider the viability of a single, independent repository, and to propose policies that guarantee that people’s online information is considered their property rather than a mere asset to social media corporations.

References

Core Aspects of Affective Metacognitive User Models

Adam Moore, Victoria Macarthur and Owen Conlan
Core Aspects of Affective Metacognitive User Models
Adam Moore¹, Victoria Macarthur¹ and Owen Conlan¹,

¹ KDEG, School of Computer Science and Statistics, Trinity College, Dublin, Republic of Ireland
{mooread, macarthv, oconlan}@scss.tcd.ie

Abstract. As User Modelling moves away from a tightly integrated adjunct of adaptive systems and into User Modelling Services provision, it is important to consider what facets or characteristics of a user might need to be contained within a user model in order to support particular functions. Here we examine previous mechanisms for creating a metacognitive user model. We then take first steps to describe the necessary characteristics of a user model we envisage being utilised by an affective metacognitive modelling service and make some suggestion for the source, form and content of such characteristics.

Keywords: Affect, metacognition, user model

1 Introduction

The successful learner has a rich cognitive repertoire of strategies and traits, which allows them to gain new knowledge, insights and understanding in a way most suited to them. Learning is not the simple transmission of information, but rather a complex process of interaction between the learner, their environment, their goals, and their informational milieu.

Technology enhanced education (eLearning) that reflects this rich learning process is an ever-evolving field. The earliest educational software of the 60’s and 70’s took a very simplistic approach, almost akin to an electronic book. With the development of hypertext systems in the 80’s-90’s Intelligent Tutoring Systems wherein a model of the learner became important, allowing them to be tailored to a greater or lesser extent to a particular type of learner or the individual. Modern systems now encompass a wide range of system architectures from mixed initiative through dialogic; serious games, inquiry-based Information Retrieval, providing animated pedagogical agents, various Virtual Learning Environments (both Open Source and Commercial) and computer supported collaborative learning.

The User Model has allowed eLearning systems to adapt to learners’ behaviour and provide adaptive feedback. The most recently developed educational software assemble interactions that infer the link between measurable outcomes (e.g. rule based inference) and resources, as well as how the user interacts with these resources. Commonly three types of knowledge are modelled to aid learning: the area being studied – the domain model, the person studying the area – the student model, and how the learning is being undertaken – the pedagogical (or androgogical) model [1].
The User Model has evolved from a component of a monolithic learning environment to become one facet of a distributed learning framework. Rather than persist the user model entirely in one application or system, they can now be delivered as a service. This means that data can be harvested from multiple sources in order to learn about the user’s collective state. In this new, distributed framework, the learning service and user model may be owned and managed independently.

As eLearning frameworks have evolved, so have their models of the learner from simple group competency-based models (e.g. stereotypes) to complex domain/skill matrices. However, many still continue to focus on the modelling of the progression of competency within a knowledge domain. The monitoring of the progression of the skills of a learner in learning is also vital, as well as the personal context of the learner. These personal traits of the learner – their affective and metacognitive states – fundamentally affect the learning process.

How do we reflect metacognitive or affective aspects of the learner in our learning systems? Metacognition involves the monitoring and subsequent regulation of cognitive processes in order to learn and solve problems [2, 3]. Metacognitive skills develop through observation of others, and subsequent internal self-monitoring [4]. Conati argues the higher the level of the user states to be captured, the more difficult they are to assess unobtrusively from simple interaction events. More fully: How do we track the individual metacognitive differences of a learner over time and discover relevant patterns of measures that can be used to predict metacognitive/affective outcomes? The next section outlines some previous systems that have attempted to do just that.

2 Metacognitive/Affective Systems

Perhaps one of the most well known early systems to model some of the above-mentioned aspects is the Cognitive Tutor – PAT (Pump Algebra Tutor) [5]. This applies the ACT-R (Adaptive Control of Thought—Rational) theory of learning and performance [6, 7]. This type of cognitive model includes procedural knowledge and declarative knowledge as well as tracing the learners’ knowledge growth over time. By developing mathematical modelling skills, learners can construct a deeper understanding of problem situations such that multiple, unanticipated questions can be addressed and answered. This has similarities with understanding of metacognitive awareness including knowledge about cognition and the regulation of cognition.

Sherlock [8], and its successor, Sherlock 2, arose out of task analysis research. These Intelligent Tutoring Systems leveraged contingent teaching that uses knowledge tracing to choose the next problem that is approximately challenging. They model the process of learning and monitor the skills not only in performing a task, but also deciding how a task should be performed. Aleven’s Help Tutor [9] supports the learner at becoming better at seeking help in geometry. The tutor keeps track of students’ knowledge growth over time using Bayesian algorithm to estimate their mastery of target skills. Although the help-seeking tutor achieved positive effects because students followed advice, they did not internalize the help-seeking principles [10].
Tutoring systems aim to emulate student-teacher interactions, however, agent based systems, such as Betty’s Brain Teachable Agent [11] emulate peer interactions. This uses AI reasoning techniques in order to externalize the thought process. Students track the agent’s metacognitive reasoning, and remediate the result if necessary. The idea comes from the fact that children can monitor errors in another person counting easier than monitoring their own errors. In the Triple-A Challenge Gameshow [12] multiple Teachable Agents, each taught by a student, compete in a game show. Students wager on whether their agents will get answer correctly. The teachable agents reason using rules taught by the students. Students showed greater motivation in learning when teaching their agents.

Narrative interaction is an important part of metacognitive skill development, whether between a student and a coach or with internal dialogues, such as learner reflection. The ACE system (Adaptive Coach for Exploration) [13] supports student exploration of mathematical functions via interactive simulations. It assesses whether a learner self-explains (metacognitive skill) their exploratory actions by using evidence for their interactions with the system and eye-tracking gaze time. Crystal Island [14] uses pedagogical agent feedback in a narrative-centred environment in order to try and keep students in an affective state that is conducive to learning. The character serves both narrative and pedagogical roles by providing task-based feedback and affective feedback. They show there was an increased performance of models including affect over those monitoring situational data alone, demonstrating the importance of empathetic support/feedback. AutoTutor [15] is a dialog-based problem-solving environment. The multimodal affect detector combines conversational cues, body language and facial features in order to infer the learner’s emotions. The face was the most indicative of the emotion, but accuracy improved using multiple indicators. Goby [16] is delivered as a separate service that is loosely coupled with the APeLS adaptive eLearning system. The metacognitive state of the learner is modeled via dialog-based interactions. The structure of Goby’s cognitive user model is analogous to that of psychometric inventories, and specifically models the MAI (Metacognitive Awareness Inventory) [17].

All of the above systems have attempted to leverage aspects of a learner’s awareness of their learning processes (metacognition) and/or their emotional (affective) state. Next is outlined an overview of mechanisms for recording and measuring these aspects.

3 Exemplar Metacognitive / Affective Models

In order to model metacognitive aspects of the learner, it seems key to represent the process or context that the metacognition arises from. Such a model is ETTHOS (Emulating Traits and Tasks in Higher Order Schema) [16], where each learner possesses Traits – these traits influence a learner’s approach to tasks. Traits are high-level metacognitive aspects such as Metacognitive Knowledge, subdivided into Factors (lower level, such as Planning). A factor can be described as a linear sum of variables. The combination of a number of related observable items describe each factor. (I pace myself while learning, I ask myself questions). The tasks are modelled as a set of activities, each activity may be broken down into Sub Activities: for example the Activity Overviewing the Learning Object (part of the “Before Starting” task), may be broken down into sub activities such as Noting important parts, Gathering information relevant to the goal, Determining what to do in detail.
Modelling the affective state of a learner is inherently problematic as it can be difficult to create an effective metric of affective states. Ortony’s Affectiv Lexicon [18] provides an often-used source of affect words grouped into affective categories. These are expansions of Ekman’s [19] basic emotions: happy, sad, anger, fear, disgust, and surprise. However, handcrafted models are difficult to generalize e.g. Dyer’s DAYDREAMER [20] – which, whilst effective in place, would be unsuitable to employ as a component of a user modelling framework. As such, the work of Liu et al [21] provides an important reference point to existing models and affective techniques. D’Mello & Graesser [15] have mapped key emotions during learning – boredom, confusion, delight, flow, frustration, and surprise.

4 Core Aspects of the Model

Given all of the above, what then, are the core aspects of an affective metacognitive user model? They can be divided into the content, form and source of the user model, as discussed below:

The content of the User Models to date that have considered either metacognitive or affective traits of the learner incorporate metrics from either structured inventories (e.g. Macarthur [16]) or bespoke solutions (e.g. Ekman [19]). The use of bespoke solutions may benefit the particular learning objectives of a course, however, if we are moving towards delivering the user model as a service, then cognitive inventories should also be considered. Over one hundred psychometric inventories are currently available for clinical, educational, and organizational evaluations. The benefit of incorporating these into the user model is that they have already been ratified and tested e.g. 16-PF [22], Myers-Briggs Type Indicator [23].

The content of the user model would therefore include, firstly, an overarching strategy for Pedagogy /Androgogy – the learning process that is being undertaken, represented by a set of formative and summative learning objectives. In particular, self-regulated learning (SRL) [24, 25] is key to learning objectives that incorporate metacognitive functions. SRL can provide a rich source of information for the user model, because the learner will engage in reflection during the SRL process. The model will also contain a narrative – that monitors and subsequently regulates communication with the learner by recording the users’ interaction with the learning environment, or through richer capture of a dialogic structure. The model should also contain aspects of cognition – the process of thought that is modelled within a metacognitive user model. Finally a learner’s emotional state must be captured, for example, by incorporating multi-dimensional axes of Ekman’s basic sextet [19].

What Form should such a User Model take? Competency-based user models have a clear metric – the comprehension of the domain in question. However, the processes discussed here are more complex. While some elements of metacognitive skills may be understood as competency based, temporal progression and, context are also important. We therefore propose a multi-dimensional matrix that records temporal, metacognitive competency and affective indices. These could be represented as both a set of metrics such as those in personality inventories as well as a number of formative learning objectives like those assessed in self-reflective journals.
The Source of the User Model can be entirely self-contained, with explicit and implicit information gathered straight from the learning environment within which it is being used. However, it could also embrace aspects of the open social web. This means that the user model content may come from a variety of sources, both purpose-built for the eLearning framework and out in the wild, such as Twitter feeds. Twitter feeds can, for example, contain affective statements, such as “I so happy I am finding my coursework very straightforward” and metacognitive information, such as, “I have spent all today planning for my tomorrow’s classes”. Information could also be taken from analysis of online forum contributions, and other social networking ephemera, such as locational and contextual cues from check-in services (e.g. Foursquare). Equally as important, from a social constructivist point of view are peer interactions through declarative living within a learner’s social graph. A rich user model comes from an in-depth inspection of the cognitive processes and affective cues collected from the user across their learning life, not just during direct encounters with learning technology. It also allows the representation of subtle affective and metacognitive characteristics, rather than simplistic steps on a chart.

4 Conclusions

As the model of learning and the learner becomes ever more complex so the need for a firm basis for the creation of a metacognitive and affective model for the learner becomes ever more necessary. We have outlined some basic characteristics we feel are key to any attempt to create such a model, based on previous work, divided into the content, form and source of the model. We suggest that such a model should be based upon externally validated inventories, with a representation of the progression of a learner through metacognitive competencies and affective states that is both temporal and stateful, respecting context. There is still much work to be done in reliably creating, updating and applying these models.

Acknowledgments. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement No. ICT 257831 (ImREAL project).

References

Extraction of Professional Interests from Social Web Profiles

Fabian Abel, Eelco Herder and Daniel Krause
Extraction of Professional Interests from Social Web Profiles

Fabian Abel¹, Eelco Herder², Daniel Krause²

¹ Web Information Systems, TU Delft, The Netherlands
f.abel@tudelft.nl
² IVS – Semantic Web Group & L3S Research Center, Leibniz University Hannover, Germany
{herder,krause}@l3s.de

Abstract. Many people share and communicate their private thoughts and opinions via systems like Facebook and Twitter. In this paper, we analyze if also professional interests of a user can be extracted from these activities and be distinguished from private interests. The results indicate that performance largely depends on the size and quality of the Social Web profiles. Methods for reducing noise and chatter for high volume profiles improve quality, but reduce diversity of the profiles.

1 Introduction

Today, one can observe a paradigm shift on the Web from a rather machine-centered view towards a more user and community-centered view. The term Social Web describes this paradigm shift and highlights a new culture of participation on the Web. As part of their daily routines, people share and communicate their thoughts and opinions via Social Web systems like Facebook³ or Twitter⁴. Therefore, a huge amount of social data is available on the Web. For example, people publish more than 65 million messages on Twitter per day⁵ about diverse topics, ranging from casual chatter to news. This large and multifaceted reservoir of social data promises benefits for various applications – in particular for systems that rely on information about their users.

Modeling interests and concerns of users has been studied already for different Social Web systems like Delicious [1], Last.fm [2] or Twitter [3]. Furthermore, there exist studies on cross-system user modeling [4] and cross-system user modeling on the Social Web in particular [5]. However, correlations between the user’s professional and private activities in the real world and her behavior in Social Web systems have not been researched extensively yet. For example, are there any correlations between a user's professional interests and her posting behavior on Twitter? Is it possible to distill a user’s professional interests from Twitter, Delicious or LinkedIn and which Social Web system is the best source to infer these interests?

³ http://facebook.com
⁴ http://twitter.com
⁵ http://techcrunch.com/2010/06/08/twitter-190-million-users/
In this paper, we answer these questions and analyze how professional interests of a user relate to her Social Web activities on Twitter, Delicious and LinkedIn. We investigate strategies for enriching the semantics of social data and study strategies for filtering social data based on interactions with other users in these systems.

2 Related Work

Any adaptation and personalization of Web systems requires user profiling, which involves the collection and interpretation of data and information about a user [6]. The Web 2.0 offers simple interfaces for Internet users to contribute user generated content – an ideal setting for performing user modeling. Especially tagging systems, such as Delicious, Flickr and Last.fm, offer the opportunity to derive tag-based user profiles. Inspired by swarm intelligence of ant colony optimization, Michlmayr et al. [1] present their Add-a-Tag approach which builds a user profile from given Delicious tags and is able to detect change of user’s interests over time and to adapt the profile accordingly. Firan et al. [2] used a tag-based user profile generated from tag-assignments of Last.fm. They were able to use the tag profile as input for search algorithms in order to generate music recommendations which outperformed classical collaborative filtering methods. In microblogging applications, like Twitter, users do not contribute keywords but short text snippets. Due to heavy use of abbreviations, noise and various languages, the semantics behind a Tweet need be captured before it can be used as input for user modeling. Therefore, Castillo et al. [7] try to classify tweets as news or chatter in order to estimate the credibility of information on Twitter. Abel et al. [3] extract topics and entities from Tweets for recommending news articles. They show that the additional processing outperforms hashtag-based profiles.

Berkovsky et al. [8] have shown that the accuracy of a user profile of a recommender system can be increased by integrating data from other recommender systems. As users often have accounts on different Web 2.0 platforms, the combination of user profiles from different platforms might increase the quality of a user profile as well. Abel et al. [5] combined form-based and tag-based Web 2.0 profiles. The tag-based profiles were extracted from Flickr, Delicious, and StumbleUpon and could not only successfully overcome the cold-start problem but also improve the quality of comprehensive single-platform-based profiles.

A remarkable observation is that the authors detected a very small overlap of the tags that a users used in different systems during the combination of the profiles. This leads to the assumption that users use different systems for different purposes: while users use Last.fm mostly to listen to music, users on Facebook connect to their friends; LinkedIn users connect to their business partners, and Twitter users might tweet about both private and business related topics. While the related work outlines that Web2.0 user profiles serve well to predict leisure interests, we are not aware of any work conducted to distinguish leisure and professional activities and predict professional interests of a user.
3 User Modeling for Mining Professional Interests

Our main goal is to extract the professional scientific interests of people from their Social Web interactions. In our first experiment, we use the scientific publications of a group of researchers as a ground truth and try to discover scientific interests of these researchers from the different Social Web platforms that they use. In the second experiment, we use these different Social Web profiles to recommend relevant publications.

3.1 Dataset

For our experiments, we used a dump of the Social Handle Archive\(^6\) (SOHARC). The dump consists of records of 99 persons, containing demographic data, contact information, and usernames of fourteen Web 2.0 platforms, like Twitter, Delicious, and LinkedIn. As the user had the choice which account information to provide, not all profiles were filled completely. In the dataset, we have 78 users who specified a Twitter account, 48 having a LinkedIn profile, 46 user with a Delicious account, and 22 users who filled all three profiles. We used the subset of 32 SOHARC users with at least two of these profiles and for whom we could retrieve the publication data.

We used Mypes\(^7\) to extract the public profile information from LinkedIn and Delicious. For LinkedIn, which provides form-based data, we used the bag-of-word approach to create user profile vectors; for Delicious, we directly used the tags to create the profile vectors. For Twitter, we crawled overall 28,293 Tweets that a) have been created by a SOHARC user, b) retweet a SOHARC user or c) mention a SOHARC user; we applied the bag-of-word approach to generate the user profiles. In a second step, we passed the Twitter profiles to OpenCalais\(^8\) and extracted entities from the Tweets for constructing entity-based Twitter profiles. Finally, we aggregated the Delicious and LinkedIn profiles with a) the bag-of-word-based Twitter profiles and b) the entity-based Twitter profiles.

To extract the professional interests of the SOHARC users, we connected them to their publication records, which are assumed to reflect their professional interests and activities. The publication records are extracted from the STELLAR Open Archive\(^9\). The publication data includes the title, abstract and keywords of publications. We related the publications to the SOHARC users by manually created mappings, which resulted in 730 assigned publications.

3.2 Overlap of Social Web Profiles with Scientific Profiles

Our main goal is to extract the professional scientific interests of people from their Social Web interactions. To gain first insights into this task, we first analyze how the terminology people use on the Social Web overlaps with the terminology of their scientific publications. In Figure 1 we therefore plot for each user $u$ the fraction of $u$’s publications that feature at least one term (excluding stopwords)
Fig. 1. Fraction of users’ publications for which terms from the title, abstract or keywords overlap with the corresponding Social Web user profile. 32 users have co-authored at least one scientific publication. The x-axis shows the $x^{th}$ user starting from 0 to 31.

in the title, abstract or keywords that also occurs in the different Social Web profiles of $u$.

The Twitter-based profile, which models a user by means of a bag of words, features much higher overlap than the LinkedIn or Delicious profiles (Figure 1). However, the entity-based Twitter user modeling - which extracts the important concepts (entities) from the bag-of-words-based Twitter - reveals that, in fact, entities overlap just a little with the entities mentioned in the title, abstract or keywords of a user’s publications. The bag-of-words-based Twitter strategy succeeds in relating 29 of 32 users to their publications, while the entity-based approach succeeds for 21 users. Furthermore, the fraction of publications per user that overlap with the user’s profile is much higher for the bag-of-words-based strategy (87%) than for the entity-based Twitter strategy (36%), Delicious strategy (38%) or LinkedIn strategy (46%).

The user modeling strategies that aggregate a user’s profile from Twitter, Delicious and LinkedIn increase the performance regarding the overlap with the user’s publications. For example, the profile aggregation strategy that combines the entity-based Twitter profiles with the corresponding Delicious and LinkedIn profiles relates, on average, 94% of the publications to the users, when examining terms that appear in both the aggregated profile and the title, abstract or keywords of a publication.

3.3 Recommending Publications based on Social Web Profiles

The positive findings on overlap between the Social Web profiles and the scientific interests of a user motivate our second experiment: now, we aim to recommend relevant publications to SOHARC users based on their Social Web profiles.

We represent each publication as a bag-of-words vector, using the title, keywords and abstract. The user profiles are generated as in our first experiment. The recommendations are based on cosine similarity between the user profile vectors and the publication vectors. As ground truth we use the co-authorship relationship.
Figure 2 shows the performance of the recommendations, as measured by the mean reciprocal rank (MRR), success at rank 1 (S@1) and precision at rank 5 and 10 (P@5 and P@10). According to the MRR, which is the inverse rank of the first hit, and S@1, Delicious outperforms LinkedIn, which on its turn outperforms Twitter. However, regarding P@10, the entity-based Twitter strategy outperforms both Delicious and LinkedIn. Our hypothesis is that Twitter profiles seem to cover more (or a broader range of) professional interests, but also seem to be much more noisy. This is supported by the fact that the semantically enriched entity-based Twitter user profiles perform better than the bag-of-word profiles. By contrast, Delicious profiles (which allow for good quality at high ranks – cf. S@1, MRR) are more focused than Twitter profiles, but they do not cover the whole variety of professional interests. This can also be observed from the drop of P@k for Delicious profiles, which is stronger than the drop for Twitter profiles.

4 Discussion and Conclusion

In this paper we discussed two experiments in which we investigated the presence of professional interests in social Web profiles and how these interests can be used for recommending publications. The results indicate that Delicious and LinkedIn profiles contain only a small amount of selected and specific interests, which cover only a small part of the professional interests. In the second experiment we have seen that these selected profiles provide qualitatively good recommendations, but fail to cover diversity. On the other hand, Twitter profiles generally contain more facets, but suffer from noise.

Obviously, there is a strong correlation between the size of the Social Web profiles and the quality of the extracted professional interest profiles ($p < 0.01$ for all individual social Web profiles and the aggregated profiles). In other words,
extraction of professional interest profiles works best for active users of social media (which also indicates that more active use typically coincides with less chatter).

We found that one way to remove chatter or noise from social Web profiles is to extract entities from the keywords. An alternative approach would be to consider only Tweets that have been retweeted by other users - assuming that users typically retweet meaningful or important messages rather than chatter. This assumption is supported by the relatively strong cosine similarity between tweet and retweet profiles (0.36) and the low similarity of retweet profiles with reply profiles (replies have been found often to be chatter, including thank-you messages). However, a preliminary evaluation showed that the potential higher quality of retweet profiles does not compensate for their significantly smaller average size (15% of the size of an average full Twitter profile).

In summary, our results confirm that professional interests can be extracted from social Web profiles. The performance of the extraction largely depends on the sizes of these profiles. Techniques such as entity extraction help to reduce the amount of noise or chatter, but this comes at the price of smaller, less diverse profiles.

References