IRIS AperTO

UNIVERSITÀ DEGLI STUDI DI TORINO

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

The Evaluation of a Social Adaptive Web Site for Cultural Events

This is the author's manuscript

Original Citation:

Availability:

This version is available http://hdl.handle.net/2318/107058 since 2020-06-29T15:37:21Z

Published version:

DOI:10.1007/s11257-012-9129-9

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)

19 November 2024

IRIS AperTO

UNIVERSITÀ DEGLI STUDI DI TORINO

This is the author's final version of the contribution published as:

Cristina Gena; Federica Cena; Fabiana Vernero; Pierluigi Grillo. The Evaluation of a Social Adaptive Web Site for Cultural Events. USER MODELING AND USER-ADAPTED INTERACTION. 23 (2-3) pp: 89-137. DOI: 10.1007/s11257-012-9129-9

The publisher's version is available at:
http://link.springer.com/content/pdf/10.1007/s11257-012-9129-9

When citing, please refer to the published version.

Link to this full text:
http://hdl.handle.net/2318/107058

This full text was downloaded from iris - AperTO: https://iris.unito.it/

iris - AperTO

University of Turin's Institutional Research Information System and Open Access Institutional Repository

Noname manuscript No.
(will be inserted by the editor)

The Evaluation of a Social Adaptive Web Site for Cultural Events

Cristina Gena, Federica Cena, Fabiana Vernero, and Pierluigi Grillo

Received: date / Accepted: date

Abstract In this paper, we present an evaluation of a social adaptive web site in the domain of cultural events, iCITY DSA, which provides information about cultural resources and events that promote the cultural heritage in the city of Turin. Using this evaluation, our objective was to investigate the actual usage of a social adaptive web site, in an effort to discover the real behavior of users, the unforeseen correlations among user actions and the consequent interactive behavior, the accuracy of both system and social recommendations and their impact on the users themselves, and the role of tagging in the user modeling process.
The major contributions of the paper are manifold: insights into user interactions with social adaptive systems; guidelines for future designs; evaluation of the tagging activity and tag meanings in relation to the application domain and thus their impact on the representation of the user model; and a demonstration of how a combination and interplay of evaluation methodologies (e.g., quantitative and qualitative) can enhance our comprehension of evaluation data.

Keywords evaluation, social adaptive system, tag-based user model, cultural events, social recommenders

1 Introduction

This paper describes the real-world evaluation of a social adaptive web site. A social adaptive web site is a social web system that provides adaptive recommendations to users. An *adaptive* system personalizes its appearance and behavior in accordance with the characteristics of the user and the context. A *social* system offers so-called "social" functionality: on the one hand, it allows users to generate and share contents; on the other, it offers them social networking facilities. Thus, a *social adaptive* system has the unique capability to exploit both the information derived from user-generated content and social networking activities

C. Gena, F. Cena, F. Vernero, P. Grillo
Dipartimento di Informatica, Università di Torino
Corso Svizzera 185; 10149 Torino, Italy
E-mail: cgena, cena, vernerof, grillo@di.unito.it

for the personalization of the user's experience and its adaptive capabilities for supporting users in their social activities.

As a test-bed for our evaluation, we chose iCITY DSA (Digital Semantic Assistant)[1], a web site that deals with cultural events (exhibitions, concerts, etc.) taking place in the city of Turin, since it is a representative example of a social adaptive web-based system. In fact, it is *social*, in that it allows its users to post their own content (e.g., insert new events and add information about them, and post comments and tags), to connect with their friends, and to create their own social networks. It is also *adaptive* in that it provides recommendations of events, based on a user's implicit and explicit interests. As a *social adaptive* system, iCITY exploits social actions, and in particular the action of tagging, in order to infer user interests and update user models (see Carmagnola [2008]for details).

The real-world evaluation of iCITY started in July 2008 and ended at the end of October (with a summer holidays break during August), and was promoted on the official web site of the municipality of Turin[2]. In total, 313 users voluntarily registered on the web site and joined in the social activity of the system. A large amount of data about user activities and opinions was then available. User actions, such as clicking on an event, a recommended event, or a tag, or adding a comment or a tag, or updating a profile or sending a message, were recorded and collected by means of *log files*, providing us with a valuable source of information, which we wanted to analyze in detail. Among all the user activities on the system, particular attention has been devoted to the analyses of *tags*, which could be either inserted by users in free text or chosen from among the system's suggestions. In particular, we analyzed both the meaning of tags and their usage. Finally, we collected information about the users' interests, users' demographics and their familiarity with technology, users' perception of the system, etc., by means of *post-usage questionnaires*.

This information allowed us to investigate the functioning of a social adaptive system. Since the beginning, in fact, iCITY was conceived as an adaptive system enriched by social components that favor user collaboration. Thus, the general goal of this real-world evaluation, as well as of previous system evaluations described in Section 4, was to discover how the social components enhance the adaptation features of the system. Our analysis concentrated mainly on: i) users' behavior in general, and ii) users' tagging activity in particular.
User behavior data were collected by means of log files; following [Gena and Weibelzahl, 2007], we exploited a log files' analysis as a kind of indirect observational method. In particular, in our analysis of log files we were inspired by "systematic observation methodology", a particular approach to quantifying behavior that is typically concerned with naturally occurring behavior observed in a real context [Bakeman and Gottman, 1997]. Using correlational analysis, we evaluated the co-occurrences and the sequences of user actions; we also applied cluster analysis. In our comparison of these elaborated data and the qualitative measures obtained from the analysis of questionnaires, we aggregated actions in order to discover the more general behavioral schemes of the users of social adaptive systems, of which iCITY is an example.
Users' tagging activity is a peculiarity of social systems. Tags characterize social systems as a bottom-up activity that offers to the user a specific view of the presented content. In a social system that is also adaptive, tags have two purposes: they can be used to model user interests and preferences, and they can be recommended to the users. The analysis of tags has concentrated on the tagging activity of users and on other tag features, such as their

[1] http://torino.mydsa.it/dsa/. Notice that this is the URL we used for the evaluation described in this paper. The web site is now available at http://www.icity.di.unito.it/dsa-dev/. An English version is also available at: http://www.icity.di.unito.it/dsa-en/.

[2] http://www.comune.torino.it/

meaning and their relation to the domain ontology, as well as their classification in more general categories.

Our results provided useful information, allowing us to understand better how adaptive and social features can benefit from each other when they co-exist in the same system, and to generalize our findings in a series of design guidelines for social adaptive systems. At a meta-level, they also confirmed the effectiveness of using a combination of evaluation methodologies. Thus, the main contributions of this paper are:

- Insights into user behavior and user interactions with social adaptive systems, and guidelines for future designs of such systems, in particular, tag-based user model systems (Section 6);
- A definition of user profiles and behavior that characterize social adaptive systems and that can be taken into account when designing this kind of system (Section 6);
- An evaluation of tagging activity and tag meanings and their role in the representation of the user model (Section 5.3). The central role of tags in iCITY is also demonstrated by the evolution of its user model toward an open user model that exports tags for interoperability purposes (Section 7);
- An analysis of user tagging behavior that can be taken into account for tagging recommendation purposes (Section 5.3);
- A demonstration of how a combination and interplay of evaluation methodologies (e.g., quantitative, qualitative) can enhance our comprehension of evaluation data. For more details see Table 8 and the related discussion.

The paper is structured as follows. Section 2 presents the state of the art of the evaluation of social adaptive systems. In Section 3, we present iCITY, the system under evaluation, focusing on its distinctive features and on the functionalities that aid our comprehension of its real-world evaluation. In Section 4, we summarize the first evaluations of iCITY, in order to establish the background for better understanding the evaluation presented in this paper. The remaining Sections describe the different steps of our real-world evaluation of iCITY. In Section 5, we present the analysis of log files containing the real usage data. We report the results of correlations, cluster analysis, precision, recall, MAE, and RMSE. Moreover, we discuss the evaluation of the meaning of tags and their impact on the user modeling process. At the end of the section, we describe the analysis of post-test questionnaires that were designed to assess user satisfaction. Section 6 provides a discussion of all the evaluations presented in Section 5. Section 7 concludes the paper and presents suggested future work, and a follow-up experiment on open user models, which we performed as a consequence of the results of this evaluation.

2 State of the Art and Related Work

The paper presents the results of an evaluation of a social adaptive web-based system. Thus, the research areas related to this work include the social adaptive web and techniques used to evaluate social adaptive systems.

2.1 Social Adaptive Web

The advent of the Social Web has radically changed the role of users, in that they have evolved from mere consumers to information producers, and more and more opportunities

are offered to users to interact with each other on the Web. In such a scenario, new challenges and opportunities for personalized systems have started to arise.

User-generated content. User-generated content (ratings, tags, comments, and so on) can be used as a source of information about a user, and exploited for adaptation and recommendation purposes, in particular Shapira et al. [2012] show how data available from social networks, specifically Facebook, can be used for the recommendation process enriching explicit user ratings and how they can play an important role in the overcome of sparsity and cold start issue of recommendation systems. In this paper we focus specifically on tags, since they are the most studied and exploited user-generated content in social systems.

Given that various studies have shown that the users' choice of tag generally reflects their interests, tags can be used to build or improve user models. Michlmayr and Cayzer [2007] show how user profiles can be built from tags by means of the Add-A-Tag algorithm, which takes into account both relationships between pairs of tags and the age of the tags. Van Setten et al. [2006] state that tags can be considered an "opinion of the annotator" and can become part of the user profile. Diederich and Iofciu [2006] use tags describing the content that is most relevant to a user to build his or her user profile; such tag-based profiles are then used to suggest publications and people with similar interests. Pirolli and Kairam [2012] propose an approach for using tags to create a learner's knowledge model. They use tags identified with expertise in a domain to identify a corpus of domain documents. Given such topical information about the domain and observed data from users, they demonstrate how to construct a model capable of inferring the users knowledge profiles across topics. Nauerz et al. [2009] introduce various tagging paradigms and explain their role in the construction of user and context models.

Similarly to such works, iCITY exploits tagging activities to determine the interests of a user. Other examples of content-based recommenders that integrate tags for modeling user interests and providing personalized ranking of items are presented in Shepitsen et al. [2008] and De Gemmis et al. [2008].

Collaborative filtering recommenders use tags in the computation of user similarity. They are based on the idea that users with similar interests have a similar tagging history, and thus, by examining the users' tagging activity, it is possible to quantify the similarity between two users [Zanardi and Capra, 2008, Nakamoto et al., 2008]. An example of a collaborative tag-based recommender is TagiCoFi (Tag informed Collaborative Filtering) [Zhen et al., 2009], a framework that includes the ratings and the tags provided by users, and builds a mathematical model for predicting user ratings based on such annotations.

Finally, addressing the issue of recommending communities of interest, Kim and El Saddik [2012] explain how various algorithms, ranging from collaborative filtering to graph-based and search-based algorithms, can be improved by including tagging information.

In the social web, a lot of user and domain data are now freely available in open formats. Many studies have appeared where user models are created that collect tags from multiple applications, since the exchange of user's tags across systems has proven to improve the quality of adaptation. For example, Abel et al. [2012] generate a user profile that aggregates user (form-based profile) and social data (tag-based profile) whose source is social networking services such as Twitter, Facebook, and LinkedIn. They introduce a service called *Mypes* that allows the integration of form-based profile as well as tag-based profile through the following actions: account mapping, profile aggregation, profile alignment, and semantic enrichment. In particular the latter point regards the meaning of tags. They cluster user tags into WordNet categories and into DBpedia URIs (for tags not contained in the WordNet dictionary). The first enrichment allows to classify the tags contained into tag-based profile into general categories of meaning such as locations, persons, animal, feeling, etc. Szomsozor

et al. [2008] present an approach that combines profiles generated on two different tagging platforms to obtain richer interest profiles. SoC-Connect [Wang et al., 2010] is a dashboard application for integrating social data from different social networks.

Social networking. Another key feature of the Social Web is the support of social relationships among users, which highlights the need for recommendation strategies that take into account "social dynamics". Information about social networks can be taken into consideration when suggesting content, friends, or groups a user could join. In iCITY, users can access the profile page of their contacts and discover what events they bookmarked or inserted. Similarly, also in iDynamicTV [Carmagnola et al., 2011a], a social adaptive system in the movie domain, the profiles of the other users can be a starting point for discovering interesting content. Other approaches use the preferences of friends to generate recommendations. Guy et al. [2009] show that recommendations derived from the target users' familiarity network (i.e., the people they actually know) achieve better performance than recommendations derived from their similarity network (i.e., unknown people with similar interests). Knowledge Sea II [Brusilovsky et al., 2004] provides social adaptive navigation support to help students find relevant items from a wide corpus of resources. While all these approaches use information from social networks to support single individuals, Loizou and Dimitrova [2012] propose an adaptation approach which takes into accout social processes to improve knowledge sharing in a virtual community, thus benefitting the community as a whole.

2.2 Evaluation approaches

The evaluation of a complex system such as a social adaptive system requires the integration of methodologies from different areas (adaptive systems, recommender systems, social sciences) and the use of both qualitative and quantitative approaches. In the following, we will present standard evaluation methodologies from these areas, making reference to their role in our evaluation approach.

2.2.1 Approaches and techniques for evaluating adaptive systems

The evaluation of an *adaptive system* requires an evaluation of the different components that collaborate to produce the adaptation. The so-called *layered approaches* have been proposed for the separate evaluation of each adaptation feature. Such approaches identify at least two layers: the content layer, and the interface layer. This idea originated from Totterdell and Boyle [1990], who first phrased the principle of layered evaluation. Karagiannidis and Sampson [2000] and Brusilovsky et al. [2001] also distinguished two levels in the adaptive process: the interaction assessment phase and the adaptation decision-making phase. Examples of layered evaluations can be found in [Brusilovsky et al., 2001, Paramythis et al., 2001, Weibelzahl, 2001, Weibelzahl and Lauer, 2001, Weibelzahl and Weber, 2001, Weibelzahl, 2003]. A more recent approach [Paramythis and Weibelzahl, 2005] identified different adaptation components and corresponding evaluation layers.

Finally, in Paramythis et al. [2010], the authors identify the following five main layers of adaptation: i) collection of input data; ii) interpretation of the collected data; iii) modeling of the current state of the "world"; iv) deciding upon adaptation; and v) applying (or instantiating) adaptation. In the evaluation of iCITY described in this paper, some of our results are explained in the light of layered evaluation, taking into account the possible effects of the interface layer on user actions and consequently on content adaptation.

According to [Gena and Weibelzahl, 2007], the evaluation process of an adaptive system can be divided into three different phases, which correspond to the typical phases of the development cycle of a system, i.e., the requirement phase, the preliminary evaluation phase, and the final evaluation phase:

- The *requirement phase* is usually the first phase in the system design process. In the case of adaptive web-based systems, it concerns the choice of the relevant features for the modeling of users (e.g., user goals and plans of the user, social and physical environment, etc.), the collection of requisites on the part of domain experts, and so on.
- The *preliminary evaluation phase* occurs during system development. It can be based on predictive or formative methods. The objectives of predictive evaluations are to make predictions, based on experts' judgment, about the performance of the interactive systems, and to prevent errors, without performing empirical evaluations together with the users. The objectives of formative evaluations are to check the preliminary design choices before actual implementation, and to obtain clues for revising the design in an iterative design-re-design process.
- The *final evaluation phase* occurs at the end of system development and its objective is to evaluate the overall quality of a system by means of users performing real tasks, for example, through usability tests, controlled experiments, and ethnographic studies.

All these evaluations were carried out for iCITY at different stages of its development cycle, as described in [Carmagnola et al., 2008] and summarized in Section 4. The real-world evaluation presented in this paper can be seen as a step in the final evaluation. Another example of a study of a similar evaluation process is reported in Zimmermann and Lorenz [2008], where two iterations of expert reviews, as well as user evaluations with questionnaires and interviews, were exploited.

The key aspects of adaptive and recommender systems that need to be evaluated are: i) whether users prefer the adaptive version of such systems to the standard one (user preferences for adaptivity) [Chin, 2001, Höök, 1997], and ii) whether adaptivity can improve system suggestions of user-relevant content (recommendation quality).

User preferences for adaptivity. This can be assessed by means of both qualitative and quantitative techniques, usually comparing the adaptive and non-adaptive versions of a system. Niu and Kay [2008] compare an adaptive and a non-adaptive version of Locator, a system that offers information about the people in a building. The study consists of a set of tasks involving each of the two versions of Locator, where subjects have to locate individuals and groups in a building. Finally, an online questionnaire is administered to users, the objective of which is to investigate their perception of the system. Moreover, the authors utilize log data and direct observation to assess time requirements and the errors that subjects make in completing the tasks.

Recommendation quality. Several metrics are used in the area of *recommender systems* in order to evaluate recommendation quality [Sarwar et al., 2001, McLaughlin and Herlocker, 2004]. In [Freyne et al., 2010], four different algorithms are proposed, which take into account the type of activity and the relationship strength among users in order to generate personalized feeds about friends' activities in a social adaptive system. The accuracy of the four proposed algorithms was evaluated by examining the ranked position, in the four alternative feed lists that were generated, of the feed items that were actually selected by users from a non-personalized feed list.

In the evaluation of iCITY described here, we focused on the assessment of recommendation quality. In fact, comparing the adaptive and non-adaptive versions of iCITY would have required us to impose some constraints on user behavior that were unlikely to occur in

real usage conditions [Gena and Weibelzahl, 2007], thus preventing us from investigating the actual behavior of users.

2.2.2 Social Science techniques

Social Sciences make use of qualitative analysis methodologies, which have proven very useful also in the area of adaptive systems[3].

In particular, since the end of the 1960s, the Grounded Theory [Strauss and Corbin, 1998] has been used in many Social Science studies to gain an understanding of complex relations between different variables, with the aim of circumventing the limits of statistical analysis alone. In fact, several studies in different fields have indicated that statistical analysis may not be sufficient or may even be misleading [Nielsen, 2004]: even when a quantitative analysis yields significant results, the actual preferences or opinions of users might remain uncaptured (Herlocker et al. [2004]). An example in the field of adaptive systems is given in Barker et al.'s study [2002], where the Grounded Theory was used to gain an understanding of the interactions that take place between learners, tutors, and the learning environment in an adaptive multimedia learning application. The aim of the study was to assess the benefit of the application to a user in terms of the delivery of effective learning. Another example is reported in Damiano et al. [2008], where the Grounded Theory is exploited to perform a qualitative analysis of the open answers gathered through a post-usage questionnaire, combining its results with those of both a quantitative analysis and a field observation. Similarly, in the evaluation of iCITY, we took inspiration from the Grounded Theory concepts in order to analyze *user questionnaires*.

Another methodology borrowed from Social Sciences is *systematic observation*, an approach to quantifying behavior [Bakeman and Gottman, 1997] that is typically concerned with naturally occurring behavior observed in a real context. As a first step, various forms of behavior are defined (behavioral codes) and then observers are asked to record whenever behaviors corresponding to such codes occur. The collected observations can be analyzed by means of non-sequential or sequential techniques. Non-sequential systematic observation is used, for instance, to answer questions about how individuals distribute their time among various activities, while sequential techniques are used to answer questions as to how behavior is sequenced in time. The results of sequential methods are more suitable for the analysis of social interaction. Systematic observation has been used to quantify behavior in Human Computer Interaction and in the adaptive web (see for instance Rizzo et al. [2005]). However, we are among the first to utilize this approach in the evaluation of social adaptive systems. This study therefore adopts the systematic observation methodology for the analysis of log files that record user behavior in the system. We then applied sequential techniques to identify recurring sequences that characterize users' interaction.

3 iCITY DSA

iCITY is a social, adaptive guide to cultural events taking place in Turin. All its users can browse the available content, which is organized and can be retrieved according to both a taxonomic and a folksonomic (i.e., tag-based) approach. Moreover, users can check event locations on a map. Only registered users, however, can enjoy the social and adaptive features of the system. In fact, they are offered personalized events lists, ordered according to

[3] For a review of the most relevant qualitative methods which can be applied to adaptive systems, see [Gena and Weibelzahl, 2007]

their preferences and contextual elements, and a sort of tag cloud displaying the usernames of other iCITY users who have similar interests. In addition, they can perform social actions (e.g., tagging or commenting on events), fill in their public profile page, create their social network indicating that other users are "friends", and use an inbox facility to send messages to their friends and receive system updates about the activities of their friends.
In the following, we will provide some information about the social aspects, user modeling, and recommendation. For more details see Carmagnola et al. [2008].

3.1 Social aspects

According to the Web 2.0 approach, a part of iCITY content is provided, via RSS feed, by the cultural portal of the municipality, TorinoCultura[4], and a part is user-generated. Notice that iCITY and TorinoCultura share a common event taxonomy (categories are: appointments, art, cinema, books, music, and theater). Moreover, each event is provided with an initial set of tags created by the automatic extraction of keywords from the event title. These tags, collected in a controlled vocabulary, are mapped to the category to which the current event belongs. Words in the controlled vocabulary and elements in the events taxonomy have a many-to-many relationship. This relationship enables the system to support the final user actively, for instance, by suggesting concepts from the vocabulary when the user is tagging an event. As far as user-generated content is concerned, registered iCITY users can post new events, add ratings, comments, tags, and further information details, as well as bookmark their favorite events.

All users have a profile page containing some information they agree to make public: personal data (age, gender, job and a free-text short description, through which users present themselves to others), current location, a list of tags they use to describe and classify interesting events, and links to the events they have posted or bookmarked.

In order to support social networking, users can define others as "friends", and then communicate with them and receive updates about their activities (e.g., whether they posted or bookmarked an event) through an inbox facility. Exploring the profile page of friends and reading update messages allows users to retrieve "social recommendations", i.e., potentially interesting events related to their friends. Moreover, a list of users who are considered to have similar interests is defined for each iCITY user, based on the idea that exploring the profile page of similar users represents another way of retrieving potentially interesting events. Similarity between pairs of users is calculated based on the following formula:

$$similarity(user_1, user_2) = \sqrt{\frac{\sum_{i=1}^{interests} (user_1[i] - user_2[i])^2}{interests}} \quad (1)$$

where $user_1$ and $user_2$ are two generic users, i represents each interest in the user model of $user_1$ and $user_2$, and *interests* is the total number of their interests. Two users are considered similar if their value of *similarity* is higher than a given threshold (in iCITY, this is currently set to 0.7 out of 1). A cloud of similar users is displayed on each page.

[4] http://www.torinocultura.it/

3.2 User Modeling and Recommendations

In iCITY, the user model is an overlay model providing, for each domain class/subclass, an estimate of the extent to which the user is interested in that feature. More specifically, it contains a probability distribution of user interests with respect to such classes. This distribution is inferred by considering the actions performed by users when they are interacting with the system. Since different actions can provide different evidence about the actual level of user interest [Kobsa et al., 2001], we assessed action informativeness and defined different weights in order to take it into account. The actions we considered are (in order of decreasing informativeness): adding events, bookmarking, tagging (namely the *action of tagging*), updating information about an event, inserting a comment/rating, visualizing an event, and clicking on a detail on the map. More specifically, a value (*newValue*) indicating user interest is computed for each class in the user model according to the following formula:

$$newValue = \frac{\sum_{i=1}^{actions} (count(i) * actionWeight)}{totalWeight} \quad (2)$$

where, for each action type *i*, *count(i)* is the total number of actions of type *i* performed by the user and *actionWeight* is the corresponding weight and *totalWeight* corresponds to the sum of all the action weights. To obtain an updated probability distribution, the new values are combined with the corresponding values in the current user models (currentValue), according to the following formula:

$$updatedValue = newValue * w1 + currentValue * w2 \quad (3)$$

where *w1* and *w2* are weights summing to 1.

Notice that, at the time of the evaluation, tags were considered as feedback only from a quantitative point of view, i.e., the number of tags was taken into account according to Formula 2, and there was no reasoning about the meaning of tags (qualitative point of view). For more details about this distinction see Section 5.

Recommendations are provided by arranging event lists in a personalized order, according to both user features (which are represented in the user model) and contextual elements, so that more relevant events are displayed at the beginning of the list. More specifically, a recommendation score is computed for each event based on a weighted mean of four criteria: interest (i.e., the level of user interest in the subcategory to which the current event belongs), position (i.e., the current user position), recency (i.e., the temporal gap between the end date of the event and the current date) and rating (i.e., the average user rating for the event). By default, at the beginning of a user interaction, the interest criterion is given the greatest weight and the other three criteria are given equal weights.
However, users can directly influence the personalization process by choosing which of the four aforementioned criteria should be given more importance when the system arranges events in order. They can select their preferred criterion either as part of their profile settings or from any site page showing a list of events. In this last case, users can immediately perceive and understand the effect of their choice, since event lists are dynamically rearranged in order to reflect the new settings. Whenever such settings are changed, the weights are rearranged so that the selected criterion is attributed the greatest weight and the other criteria are given equal weights. Let us suppose a user, who is particularly interested in cinema, logs

into iCITY. Since "interest" is the most important criterion as a default, events in the "cinema" category will be presented in the top positions in iCITY event lists, which are ordered taking into account also their mean rating, position, and recency. If the user decides to set "rating" as the most important criterion, cinema-related events with a low mean rating will probably disappear from the top positions, while events with a high mean rating in other categories will take their place.

In order to explain system recommendations to users, recommended events are emphasized with visual cues, according to the "adaptive annotation technique" [Brusilovsky, 1996]. According to this technique, links can be annotated with icons or other cues (e.g., colors) in order to help the user in the selection of the most relevant suggested items. In iCITY, colored thumbs-up icons are used to indicate the predicted level of interest for a certain user. Conversely, stars are used to represent graphically the mean user ratings given by the community of users. The choice of these particular icons was inspired by the work described in Cena et al. [2005], wherein the authors describe an evaluation of adaptive annotation techniques. According to the results that were reported, users associated stars with a general qualitative judgment, while they related emoticons to feelings expressed by the system itself. Adhering to these findings, we opted for stars to communicate general messages, while we decided to utilize thumbs-up icons to communicate personalized messages.
Finally, iCITY users can understand recommendations better and inspect the system assumptions about their interests by accessing their *open user model* [Kay, 2006], which was simply displayed as a table summarizing their levels of interest, as they were inferred by the system, with respect to each category in the event taxonomy (see Figure 6). However, users are not yet allowed to modify their model (see Section 7 for details on the open user model).

4 The Past Evaluations of iCITY

We start by briefly reporting the first evaluations of iCITY that we had carried out previously; for more details see [Carmagnola et al., 2008]. Following the steps proposed by Gena and Weibelzahl [2007], we carried out different evaluations at different stages of development: the *requirement phase*, the *preliminary evaluation phase*, and the *final evaluation phase* (see Section 2).

In the *requirement phase*, the objective of our evaluation was to gather requirements for exploring how users tag information. We chose a list of events from the RSS channel that feeds iCITY and then we asked 39 users to tag the events or to choose from the description of the resources. We collected 217 tags and analyzed them inductively, according to the principles of the Grounded Theory [Strauss and Corbin, 1998] with the goal of defining the main classification categories. The first two categories (and their corresponding frequencies) that we considered are: *i)* proposed tags (tags derived from the resource description) (76%); *ii)* free-text tags (tags directly inserted by users) (24%). Taking into account other properties related to the tagged resource, other sub-categories emerged, as can be seen in Table 1.

In the *preliminary evaluation phase*, during the development of iCITY, we carried out two different evaluations: a heuristic evaluation, performed by an HCI expert and an adaptive web expert, and two sets of usability tests of the scenario-based prototype of the system. For the former, the experts were asked to follow Jameson's five usability challenges [Jameson, 2003] for adaptive interfaces; while for the latter, we designed two different sets of tasks in order to guarantee that all the features were evaluated. These preliminary evaluations led to a re-design of both the user interface (e.g., labeling, shortened long event descriptions, addition of some previously lacking feedback messages, etc.) and some aspects of the sys-

Category	Description	Percentage
Specific tags	Tags that add some specification about the resource (e.g., Bono)	61.19%
Generic tags	Tags that classify the resource in a more general way (e.g., concert)	22.37%
Contextual tags	Tags about the context of the resource: location, time, etc. (e.g., Turin, August)	13.24%
Synonym tags	Tags that are synonyms of terms in the resource description (e.g., live, show)	2.74%
Unknown tags	Unknown words, e.g., unhyphenated compound words such as "Turin-Concerts"	2.17%

Table 1 Tag categories emerged from requirement-phase evaluation and tag frequencies. The reported examples refer to the event "U2 360° Tour 2010 in Turin"

Category	Frequency	Percentage
Generic tags	212	49%
Specific tags	114	26%
Spatial tags	58	13%
Unknown tags	28	6%
Subjective tags	19	4%
Synonym tags	6	1%

Table 2 Tag frequencies that emerged from final evaluation

tem functionalities (e.g., the rationale for recommendations was made clear, and also the user model was made scrutable, the user was given control of recommendation sorting, the user's bookmarked events were made public, etc.). The details are extensively discussed in [Carmagnola et al., 2008].

For the *final evaluation phase*, we decided to test the system under real conditions with users performing real tasks. We selected a group of 20 users, all target users of the system. We asked them to register onto the system and to use it every day for two weeks. After this period, they compiled a free report containing the problems they experienced, and they were also asked to evaluate their read-only open user model that contained the scores that the system assigned to each category of interest. Eight of the 20 users said the user model was correct, while 12 subjects subjectively re-assigned the values of some of their categories of interest. Concerning the accuracy of the system predictions, we obtained a medium MAE of 0.11 in a range from 0 to 1, which according to the literature can be considered as a good value[5].

We also examined all the tags inserted by the users. We found 437 tags used to annotate a total of 183 different events. We manually analyzed the meaning of the tags by classifying them into the categories described above. Of all the tags, 321 (73%) were *free-text tags*, i.e., tags directly inserted by users, and 116 (27%) were *proposed tags*, i.e., tags proposed by the system. The results for the other categories are presented in Table 2.

In order to investigate further the role of tags in the definition of the user model in iCITY, and their impact on the accuracy of recommendations, we carried out a second evaluation. We involved the same users as in the first evaluation and asked them to use iCITY for two weeks, every day. However, this time they were explicitly asked not to use any tags. After the experimental period, they were again required to evaluate their open user model. This time we obtained a medium MAE of 0.40 in a range from 0 to 1. This value was higher than

[5] Good et al. [1999] suggest that good values of MAE should be near to 0.7, on a scale of values ranging from 0 to 5.

the previous one (0.11) and this demonstrates that, in iCITY, tags have an important role in the definition of the user model and the accuracy of the recommendations.

5 The real-world evaluation

Having analyzed the results of the evaluation described above, we decided to organize a real-world evaluation, in order not only to extend the validity of our past results but also to investigate the real usage of the system, especially as far as its social and adaptation features are concerned. Data collected when a large number of real users do real tasks in a real context of use are very relevant to the evaluation of adaptive systems. Many authors have emphasized the importance of studies involving real users in the real context of usage as well as the lack of case studies reported in the literature [Chin, 2001, Weibelzahl, 2003, Gena, 2005]. Moreover the analysis of the real behavior of the users under real conditions makes it possible to analyze the facts more in depth rather than under experimental conditions, wherein the ultimate goal is to explain the cause–effect relationships between variables and make generalizations on the basis of the obtained results.

The real-world evaluation of iCITY started in July 2008 and ended at the end of October. During this period, the system and the experimentation had been promoted on the official web site of the municipality of Turin[6] with a banner inserted in its home page. Starting from this time, the web site of iCITY was voluntarily visited by an average of 41 visitors per day, including registered users.

Google Analytics[7], which was used to monitor from where the users came and their general behavior on the web site, reported the following data:

- Visits: 5,464;
- Absolute Unique Visitors: 2,989;
- Pageviews: 45,195;
- Average Pageviews: 8.27;
- Time on Site: 5,52 minutes;
- New Visits: 54.47 %.

During the trial, the pages of this site were viewed a total of 45,195 times. In particular, the pages that were viewed most often were:

- Home page: 13,418 times;
- Last events: 1,437 times;
- Online user: 655 times;
- Registration: 604 times;
- Help: 369 times.

The main goals of the real-world evaluation were:

1. to analyze the real user behavior (browsing, selecting, tagging, rating, commenting, etc.) when interacting with a social adaptive system in order to gather insights relevant to future re-design and to be able to formulate guidelines for the design of such a type of system;
2. to evaluate the users' selection of system recommendations and social recommendations and their accuracy by comparing the system's assumptions with real user preferences;

[6] http://www.comune.torino.it/

[7] http://www.google.com/analytics/it-IT/

3. to evaluate the tagging activity of users. In particular, we were interested in: i) discovering how the meaning of tags is related to the user interests in the user model; ii) classifying users' tags in the general-purpose classification described in Section 4 in order to obtain an empirical confirmation of the correctness of this classification and to inspire the design of a tag-based user model and more general systems recommending tags;
4. to investigate final users' satisfaction with the system and their opinions of it.

In particular, we achieved goals 1 and 2 by means of the log-based field study. Concerning the metrics, for the evaluation of the behavior of the user (goal 1), we decided to use statistical correlations and cluster analysis in a systematic observation perspective in order to analyze the co-occurrence of behavioral code units and their frequent sequences, so that interaction codes could emerge. In the evaluation of the content selection process (goal 2), we exploited standard measures, such as precision and recall, MAE, and RMSE to evaluate the accuracy of both the adapted and social content.
As far as the tagging activity of users (goal 3) is concerned, we repeated the analysis described in Section 4, by classifying manually the tags that users generated during the trial with respect to the categories that had emerged from the previous analysis. Moreover, in order to assess the impact of tags on user models, we exploited a semi-automatic component that classified user tags with respect to the system domain taxonomy. We calculated user interests with respect to each class in the taxonomy, based on the number of tags that were mapped to that class, and compared such interest values with i) the corresponding values in the current user models, and ii) interest values that users explicitly declared.
Finally, to achieve goal 4, our investigation was based on an analysis of explicit users' answers, which were collected by means of an online post-usage questionnaire. We investigated user satisfaction based on user self-evaluation of their overall satisfaction with the system, as well as on their assessments of specific HCI aspects, such as meaning of icons, perception of recommendations, and so on.
The following subsections present in detail how we attained these goals.

5.1 The analysis of real users behavior

Subjects. 313 users out of 2,989 (10.47%) unique visitors voluntarily logged onto the web site. They filled in a form that collected their socio-demographic data. They were prevalently male (60.67%), while females constituted 39.33%, with an age distribution ranging from 21 to 65 years-old, and an average age of approximately 34 years.
97 users out of 313 (31%) registered onto the system, visited the web site, and performed actions. Some of them updated their personal profile (sometimes in a very detailed way) – and then never logged back on. The remaining 216 (69%) were returning visitors. We calculated that these 216 users logged on an average of 9.3 times. Note that for the study reported in this paper, we have analyzed the activities of all these 313 "registered users".

Measures. The activity of "registered users" was recorded by means of log files collected from the web server. In particular, all page requests and all user actions (updates, content insertion or editing, etc.) were logged. In order to aid the interpretations of the results, we devised a classification of user actions to guide the analysis of the results. Note that the more frequent actions are shown in Figure 1. We divided the actions into web actions, social web actions, and adaptive web actions. An action is classified in a particular group if it is typical of the corresponding type of web site (i.e., regular/adaptive/social). For example, clicking on a tag usually occurs in web sites that take into consideration the keywords (tags) used by

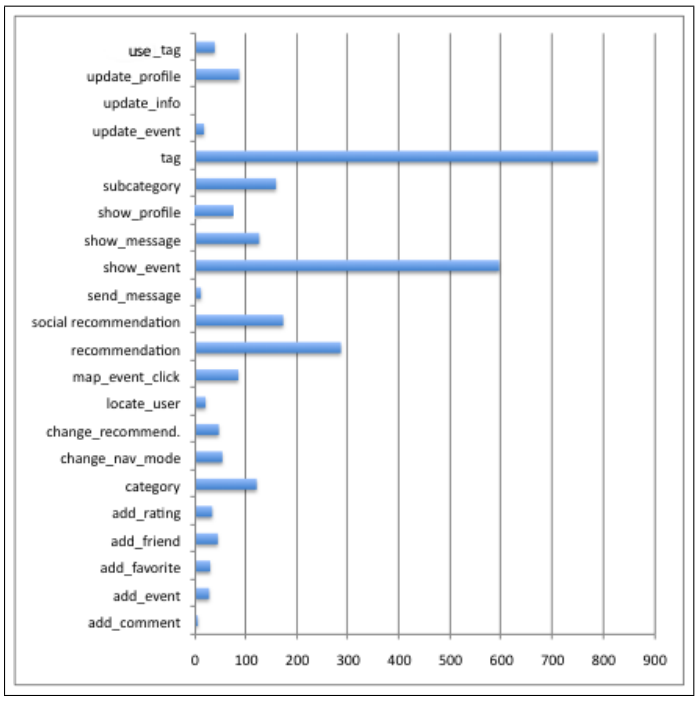

Fig. 1 The most frequent actions registered by the logging component of the system. The Y axis constitutes the actions, while the X axis constitutes frequency.

the community of users. In this perspective, this action is classified as a social web action. This classification i) had the aim of aiding the interpretation of the results and in particular of discovering the possible interplay between different kinds of actions, ii) does not consider the impact of actions on adaptation, iii) excludes those actions never or almost never used, such as "remove rating", "remove favorite", and so on.
Web Actions are typical actions on web sites, such as :

- clicking on an event (show_event). This action includes all the clicks made on an event title to access more information about the event. Note that all the clicks made on an event title are classified under this label, comprising the click on a recommended event and the click on a "socially recommended event", which are described below. This is because the logging component cross-classifies the clicks among these categories. However, we will take into account this aspect in the further analysis;
- clicking on a category (category), namely clicking on a global category of navigation;
- clicking on a subcategory (subcategory), namely clicking on a local category of navigation;
- clicking on the map (map_event_click), namely clicking on the map to see a localized event;
- changing navigation mode (change_nav_mode), namely changing between textual and geo-visual navigation modes.

Social Web Actions are those actions typical of communities, social networks, and social web sites:

- adding comments to an event (add_comment);

- adding a new event (add_event);
- adding a friend (add_friend);
- adding tags to an event (tag);
- clicking on a tag to see the related events (show_tag_event);
- clicking on a social recommendation (recommendation_social), namely clicking on a favorite event of a friend;
- clicking on a user profile (show_profile), namely clicking on the profile section of a friend or other user;
- updating an event (update_event);
- updating their profile (update_profile);
- adding a favorite event (add_favourite), namely adding an event to the list of favorite events (bookmarks)*;
- rating an event (add_rating), namely clicking on the 4 stars near the event title to rate an event *;
- sending a message to a friend of the community (send_message);
- open a message received from the system about a friend of the community (show_message);
- locating a user position on the map (locate_user).

Adaptive Web Actions are actions that are exclusive to adaptive/recommender web sites:

- changing the recommendation criteria (change_recommend), namely changing the ranking modality of an event list, e.g., interest, recency, average rating of other users, proximity;
- clicking on a recommended event (recommendation), namely clicking on a recommended event.

* These actions belong to both the social web and the adaptive web. Since such actions are bottom-up, close to the Web 2.0 philosophy, we consider them mainly as social actions.

Results. To analyze the real user behavior when interacting with a social adaptive system (goal 1), we correlated all the logged user actions (detailed in the lists above) made throughout the evaluation period, which, as stated in [Gena and Weibelzahl, 2007], can be considered a kind of indirect observation of user behavior, and we systematically analyzed them. In the literature, systematic observation is defined as a particular approach to quantifying behavior. This approach is typically concerned with naturally occurring behavior observed in a real context [Bakeman and Gottman, 1997]. The aim of systematic observation is primarily to define various forms of behavior (behavioral codes); observers are then asked to record whenever behavior corresponding to the predefined codes occurs. The behavioral code units are all the possible actions the users can perform, which we have considered as independent variables. In order to discover significant relationships between all the variables and co-occurrences, we measured Pearson correlation, since scores showed a normal distribution (see [Keppel et al., 1998] for details). As for interaction codes, they will inductively emerge when we put together these results with a cluster analysis and the qualitative data from the questionnaires. They will be detailed in Section 6. All the significant correlations that we found are reported in *Appendix 1*.
In order to interpret the correlational data, some clarification is needed. First, some correlations between clicking on an event and other actions may be not very relevant. These are the correlations highlighted in italics in *Appendix 1*. For instance, due to the interaction design of the interface, a user, before rating an event/adding the event as favourite/adding tags, has to click on the event. These actions have a high correlation with the action of clicking due to the interface's constraint (see Figure 2 and the italic correlations in the list). This is a

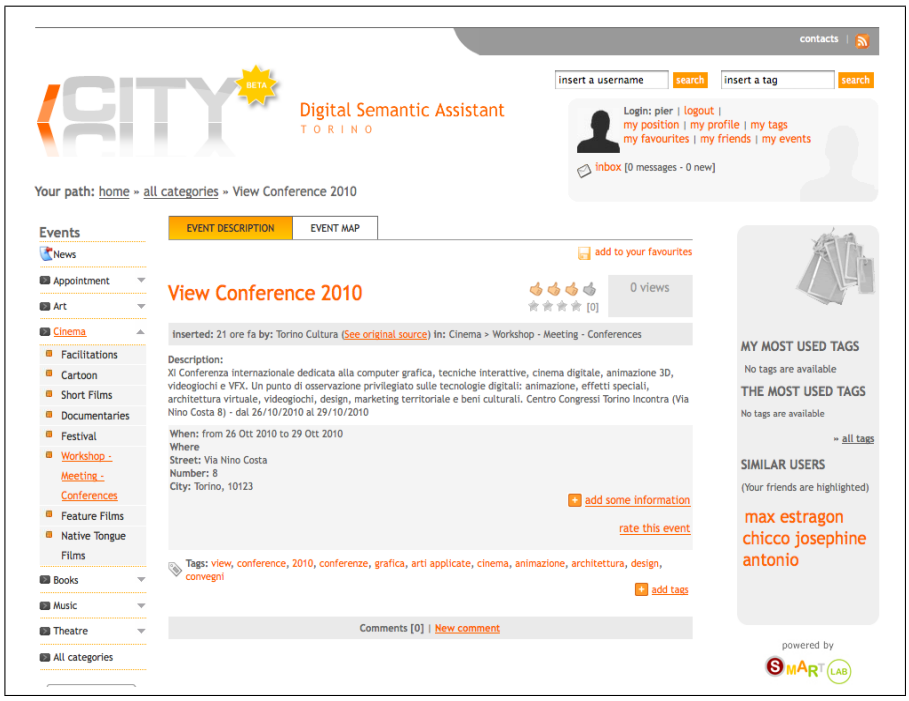

Fig. 2 The consequence of having clicked on an event

typical example of a layered approach to the evaluation that takes into account the effect of interface constraints on the interpretation of user actions. Thus, the design of the interface and its constraints render this correlation inherent or meaningless. However, it is interesting to model a sequential analysis of these actions, i.e., how many times the users performed those actions after having clicked on an event. 6% of actions of clicking were followed by the action of rating, while 5% of actions of clicking were followed by the action of adding the same event as a favourite (r=0.863, significant at the 0.01 level). Concerning the adding of tags after having clicked on a event, *60% of users tagged an event*, and every event received an average of 1.32 tags. Secondly, we discarded i) the correlations where there was not enough usage to justify a conclusion, e.g., those involving adding a comment (5 actions, 0.17% of total actions) and updating personal information (6 actions, 0.21%), and so on, since these actions happened very infrequently. However, we did not discard some more frequent actions, such as adding a favourite event (29 actions, 0.99%), rating an event (33 actions, 1.12%), and adding an event (27 actions, 0.92%), due to their relevance as social actions; ii) the meaningless negative correlations, e.g., changing recommendation criteria and opening a message; iii) correlations for which we found no interesting interpretation, involving, e.g., actions that do not have any kind of correlation such as clicking on a event and localizing a user position, clicking on an event and opening a message, and so on. The most interesting correlations occurred between either two social web actions, or a social web action and an adaptive web action. With the objective of identifying recurring interaction codes in user behavior, we performed a sequential analysis in order to determine, for each selected correlation, how often the two involved actions occurred consecutively on the

same event. Notice that, at this stage of our analysis, we focused on the action type level, in order to describe general phenomena.

Correlations between social web actions. Users who open a message to get information about their friends' activities always click on events (r=1.00, significant at the 0.01 level). 58% of these combined actions happened one before the other, namely, the click after having opened the message. These messages were all advising the user that a friend had either rated or added an event as favourite.
Users who add a favourite event frequently rate an event, (r=0.934, significant at the 0.01 level); 18% of these combined actions occur together, related to the same event. Users who add favourites, also click on tags to navigate (r=1.00, significant at the 0.01 level), and 9% of actions occur together.
Users who add a tag frequently both rate an event (r=0.890, significant at the 0.01 level) and add a favourite (r=0.884, significant at the 0.01 level); 9% of the actions of rating and tagging occur together on the the same event, and 16% of actions of adding a favourite and tagging occur together and relate to the same event.
Users who add a favourite click quite often on social recommendations (r= 0.750 significant at the 0.01 level), while users adding tags click on a social recommendation not so frequently (r= 0.570 significant at the 0.01 level). However, the former two actions occurred together on the same event 11% of the time, while the latter two 31% of the time. We also found some negative correlations that seem to underline an inverse trend between single click actions and text insertion actions: rating an event and updating an event (r=-1.00, significant at the 0.01 level); clicking on the map and adding an event (r=-1.00, significant at the 0.01 level); adding a favourite event and updating the profile (r=-1.00, significant at the 0.01 level). These correlations suggest that users can play different roles: some seem to be more interested in *consuming content* and *performing few demanding actions*, which are related to content evaluation and organization, while others seem to be more interested in *generating new content* rather than in consuming the existing content.

Correlations between adaptive web actions and social web actions. We found that the action of clicking on a recommended event is strongly related to the action of rating an event (r=0.929, significant at the 0.01 level). Considering all users who rated at least one event and visualized at least one recommended event, users rated a recommended event after visualizing it in 7% of cases. Also the actions of adding a favourite and clicking on a recommended event are quite strongly related (r=0.866, significant at the 0.01 level). Considering all users who added at least one event to their favourites and visualized at least one recommended event, we found that, *in 72% of cases, users actually bookmarked a recommended event after visualizing it*. Clicking on a recommended event is also quite often correlated to adding a tag (r= 0.773 significant at the 0.01 level). Considering all users who added tags to at least one event and visualized at least one recommended event, we found that *they added tags after clicking on a recommended event in 12% of cases*. Finally, clicking on a recommended event and clicking on a social recommendation (r=0.470, significant at the 0.01 level) occur together infrequently. This suggests that *different types of users may prefer different types of recommendations*, either suggested by other users or generated by the system.

Figures 3 and 4, and the clustering results discussed below, summarize our main findings and present the interaction codes that emerged from the sequential analysis of user behavior. Figure 3 presents the main interaction codes in user behavior that we have identified with respect to social web actions. We found that more than half the users who open a message containing a reference to an event immediately visualize the event itself, allowing us to conclude that personal messages are a good prompt to immediate content fruition. Moreover, we found that users who visualize a socially recommended event add tags to it more often

Fig. 3 Example of frequent interaction codes

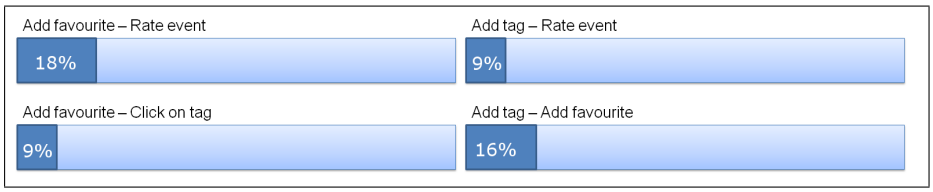

Fig. 4 Co-occurrence of social actions on the same event

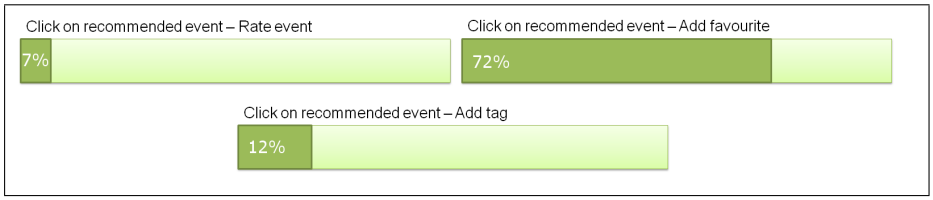

Fig. 5 Co-occurrence of social and adaptive actions on the same event

than they bookmark it (33% vs 11%); on the whole, *users either bookmark or tag a socially recommended event in 44% of cases after viewing it*.
Figure 4 visually summarizes the co-occurrences of social actions on a given event. In all cases, two social actions occur on the same event in less than 20% of the cases. The most significant co-occurrences involve the "add tag-add favourite" and "add rating-add favourite" pairs of events. This suggests that users who are very interested in an event and bookmark it are also quite willing to annotate it.
In Figure 5, we summarize the co-occurrences of social and adaptive actions on a given event. We can note that cases when a user visualizes a recommended event and then also bookmarks it represent 72% of all cases where the same user has both clicked on a recommended event and bookmarked it.

To achieve a more detailed comprehension of the user behavior, we also performed a TwoStep Cluster Analysis to sketch some emerging user profiles. The TwoStep Cluster Analysis procedure is an exploratory tool designed to reveal natural groupings (or clusters) within a dataset that would otherwise not be apparent. This clustering was derived automatically with SPSS software[8]. Three clusters emerged:

- *Heavy users cluster* (46% of cases) is characterized by the action show_event (55.6%). The main other actions are: update_profile, add_friend, map_event_click, change_nav_mode, add_rating, and send_message. This cluster identifies "heavy users" who exploit the main features of the systems, especially the social ones;
- *Email users cluster* (32.9% of cases) is characterized by the action log_out (32.9%). The main other actions are: remove_message, show_message, click_tags, and add_favourite.

[8] http://www.spss.com/

This cluster identifies "email users" who access the web site using the email messages sent by the web site as a starting point of navigation;
- *Single log-in users cluster* (26.1% of cases) is characterized by the action log_in (60.7%). The main other actions are: registration, locate_user. This cluster identifies "single log-in users", those who register and never log back in.

To summarize, Cluster 1 identifies "*heavy users*", socially-oriented users performing particular actions that involve, directly or indirectly, friends. Cluster 2 identifies "*email users*" who are interested in the messages they receive from the system, and thus they are interested in the actions of their friends. Cluster 3 is not particularly interesting for our analysis. We observe that actions, such as following social recommendations starting from the profile page of other users, following system recommendations and adding tags, even if very frequent, do not aggregate users showing similar behavior and thus do not characterize any cluster.

5.2 The evaluation of the user's selection process and the accuracy of system and social recommendations

To evaluate the *user's selection of the adapted contents* (goal 2), we first analyzed precision and recall (for details see [Herlocker et al., 2004]). Notice that, for the calculation of these metrics, we considered the click on a link, specifically the click on the title of an event in order to obtain more information about it, as a selective user action[9] [Kobsa et al., 2001]. Thus, we considered a click on an event title as an indicator of user interest, and clicked events have been considered as user-relevant content in the calculation of both the *precision* (the ratio between the recommended user-relevant contents and all the contents recommended to the user, see Formula 4) and the *recall* (the ratio between the recommended user-relevant contents and all user-relevant contents present in the content collection, thus also including the content that the system does not suggest, even if it can be relevant to the user, see Formula 5).

$$precision_{SystemRecommendations} = \frac{|clicked\_recommended\_events|}{|recommended\_events|} \tag{4}$$

$$recall_{SystemRecommendations} = \frac{|clicked\_recommended\_events|}{|clicked\_events|} \tag{5}$$

The calculated values were only partially satisfying: 0.74 for precision and 0.4 for recall. A reason for this could be that in the calculation of precision and recall we considered the actions of all registered users, including the 97 (31%) single-visit users. In this case, as well as for users who logged in less than five times, the recommendations are generic and based on the event's popularity; they are not filtered for specific users, since the update of the user model occurs after every five interactive sessions. Moreover the system has been designed explicitly to favour precision, filtering proactively user-tailored contents, instead of recall, somehow hiding the contents supposed to be not relevant for the user (notice that usually precision and recall show an inverse behavior). We also analyzed precision and recall without considering the 97 (31%) single-visit users. The results are more satisfying: 0.85 for precision and 0.43 for recall.

[9] According to Kobsa at al. [2001], the most frequent kind of interaction with web-based systems is clicking on a link. Such a selective action can be regarded as an indicator for several types of user data: interest, unfamiliarity, and preferences.

We then investigated *user's selection of social recommendations* (goal 3). By social recommendations, we mean the preferred items of the user's friends in the community. We have therefore considered both: i) the action of clicking for more information on a link presented in the bookmarked events section of a friend; ii) the follow-up after the action of opening a message received from the system about what action a friend has recently performed, such as adding a bookmark, rating an event, tagging an event or posting a new event. Thus, if the user clicks on the favourite event described in the message, this action is considered in the calculation of the metrics. For this group of actions, which we called social recommendations and indeed express the contents that friends like, we also calculated precision and recall. Notice that, in this case, clicking on the name of a friend's (bookmarked/rated) event in order to obtain more information about the event was considered to be a selective user action. For this group of actions, we also calculated *precision* (the ratio between user clicks on the events their friends like and all the content that their friends like, see Formula 6) and *recall* (the ratio between user clicks on the events their friends like and all user clicks on events present in the content collection, see Formula 7).

$$precision_{SocialRecommendations} = \frac{|clicked\_socially\_recommended\_events|}{|socially\_recommended\_events|} \tag{6}$$

$$recall_{SocialRecommendations} = \frac{|clicked\_socially\_recommended\_events|}{|clicked\_events|} \tag{7}$$

The results were partially satisfying: 0.45 for precision and 0.24 for recall. However, this is not surprising since the interface promotes the system recommendations (e.g., in the home page, in the latest events page, at the beginning of every sections, etc.), which are thus more likely to be followed than social recommendations. Social recommendations are listed in the friend profile section, and the interested user therefore has to look for them actively or read a message in the Inbox section that gives information about the favourite events of a friend. Note that this is another example of the layered approach to the evaluation: interface constraints bias the negative results obtained by social recommendations, which might perform better if promoted differently. Moreover we have to note that the 97 single-visit users are automatically excluded from social recommendations since they did not select any friends.

The results for recommendations and social recommendations are also confirmed by the correlational data described below. More specifically, clicking on an event and clicking on a recommended event are actions that are quite often correlated (r=0.850, significant at the 0.01 level), while clicking on an event and clicking on a social recommendation are actions that are not so strongly correlated (r=0.440, significant at the 0.01 level) (see Figure 1). For the sake of clarity, we should remember that the action of "clicking on an event" also comprises clicking on events that have been recommended either by the system or by friends. Thus, these latter two categories of actions are a part of the general action "clicking on an event" (show_event), and the correlations between these actions emphasize only how many clicks on an event have been recommended by the system or by friends.

In order to assess the *accuracy of recommendations* (goal 2), we based our investigation on both user answers to an explicit question (*perceived recommendation accuracy*) in the questionnaire described in Section 5.4, and objective accuracy metrics (*objective recommendation accuracy*). Note that the perceived recommendation accuracy is intended to show how users perceived system recommendations, namely matching their interests or not, while the objective recommendation accuracy is calculated by statistical accuracy metrics.

Concerning perceived recommendation accuracy, 16% of users said iCITY had never suggested events matching their interests, 49% said sometimes, 34% said often, and 1% always. Concerning objective recommendation accuracy, we have exploited two well known accuracy metrics: the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE). MAE measures the average absolute deviation between a predicted rating and the user's true rating for that value. RMSE, which squares the errors before summing them, puts more emphasis on large errors [Herlocker et al., 2004]. Users were asked to express a rating for a set of cultural interests such as "Appointments", "Art", "Cinema", "Books", "Music", "Theater", by assessing them on a scale ranging from 0 to 10, where 0 indicates minimum interest and 10 indicates maximum interest. Note that we asked users to rate in a scale ranging from 0 to 10 since the user model represents the user-inferred interests on a range from 0 to 1. Thus, the 0-10 granularity permits a more correct estimation of the deviation between predicted ratings and user ratings. These interests, which represent the domain categories of iCITY, overlay the user model of the system. We have compared these explicit ratings to the system's predicted ratings for each user in order to calculate the accuracy of the predictions. The obtained values were, respectively: a medium MAE of 0.06, and a medium RMSE of 0.07, which we considered very satisfying. It should be noted that, both for every user's ratings and for system predictions, the scores assigned to each category of interest were normalized to sum up to 1. Therefore the value of both MAE and RMSE refer to these normalized values.
On the whole, the results of this analysis revealed that the system quite accurately infers user interests, which confirms and explains the relatively good result we obtained for recommendation precision. Considering all these measures, as well as user opinions, we can conclude that recommendations performed quite well, and most users were aware of the recommendation features. Social recommendations performed worse, but were probably penalized by the interface constraints.

5.3 The evaluation of user tagging

As described in Section 3, iCITY provides recommendations on cultural events, based on the user model enriched with tags that we exploited to infer user features. Our initial hypothesis was that tags can be useful for increasing the system knowledge about users, since from tags we can infer a user's interests. The tagging activity can be analyzed from a quantitative and a qualitative point of view. The quantitative point of view considers the *action of tagging*, while the qualitative point of view regards the *meaning of tags* with which the user annotates the resources. The quantitative point of view was taken into account in the first release of iCITY (see Section 3) and in its past evaluations (see Section 4). In this new evaluation of user tagging, we focused on the qualitative point of view. In the following, we first clarify these two perspectives, and then present the evaluation of the collected tags.

The quantitative point of view. According to the *quantitative point of view*, tagging, and, more generally, annotating are considered as possible actions a user can perform on a social web site. Like other kinds of *usage data* [Kobsa et al., 2001] such as clicking, buying, etc., these actions represent important feedback from the user. In fact, users can tag with different purposes: to categorize a resource for the community, or to describe it for future retrieval, or to express an opinion [Marlow et al., 2006], etc.. In our case, we refer to tagging for future retrieval, i.e., labeling a resource for the purpose of finding it later. As described in 2, the action of tagging is a stronger indicator of user interests [Kobsa et al., 2001] than simply clicking on a link, and therefore should be analyzed in order to draw interesting inferences

about the user model. At the time of the first iCITY evaluation (for details see Carmagnola et al. [2008]) tags were considered only in a quantitative way by the user modeling component, because the system considered only the "action of tagging" and did not analyze the meaning of tags.

The qualitative point of view. This investigates the possibility of reasoning about the meaning of tags in order to infer knowledge about the users. We decided to analyze the tag semantics by exploiting: i) an ontology of the application domain, including relevant concepts and the relationship of the concepts, and ii) a database that specifies a list of terms for each concept in the ontology, in the form of a lexical database such as WordNet[10]. For this evaluation, we developed an automatic component that is responsible for analyzing the meaning of the tags, looking for correspondences between the tags and the *synsets* and the *domains* of the MultiWordNet database[11]. Note that the classes and subclasses of the iCITY event ontology are mapped on the corresponding synset and domains as a semantic enrichment step. These relations are in "one to many" cardinality, since a class/subclass of iCITY may correspond to one or more synset/domain of MultiWordNet. If one or more correspondences between tags and synsets/domain are found, tags are linked to the class/subclass of the user model interests (which overlay on the ontology domain) and thus are considered as an indicator of user interests for that class/subclass.

5.3.1 The evaluation of the meaning of tags

As described above, for the construction of the tag-based user model, we have assumed that tags, and their meanings, may somehow reveal a level of interest of the user in the meaning of those tags. For instance, a user interested in music could save as favorites a lot of music events, tagging them with the keywords "live music" and "concert". Thus a user modeling system could infer a user's interest in live music and concerts, and more generally in the concept of music.
During the real-world evaluation, we collected a number of tags (namely, 788, see Figure 1). We used them to carry out an evaluation of the impact of tags with respect to the user model. In particular, we wanted to discover whether the meaning of the tags is really related to the user interests in the user model, in particular to the classes/subclasses of user interests in the user model.
Note that we considered the tags inserted by users whose user models contain updated values. As explained in Section 3, the model is initialized to a uniform probability distribution representing the same interest value for each domain class. After a certain number of interactions, such values are updated in order to reflect the user interests better.

To analyze the meaning of the tags, we exploited the automatic component described above. Moreover, we manually refined the automatic classification every time a tag was discarded, and we checked whether the tags were correctly dropped. For instance, in a few cases, users tagged the events using adjectives, such as "modern" or "contemporary", which were automatically discarded from the system. We manually re-classified them in the class/subclass to which they referred (i.e., music and arts). Major problems of disambiguation were found in the "Appointment" class, since this was a kind of "umbrella" class with few tags related to it. For instance, appointments tagged as "photography", "conference", "concert", "exhibition", and so on were not automatically classified in that class even

[10] http://wordnet.princeton.edu/

[11] In MultiWordNet (http://multiwordnet.itc.it/), and WordNet, each synset is annotated with at least one domain label, selected from a set of about two hundred labels that constitute the so-called WordNet Domains.

Class	Tag Frequency	Event frequency
"Appointment"	53 (13.18%)	175 (13.82%)
"Art"	65 (16.17%)	84 (6.64%)
"Cinema"	32 (7.96%)	173 (13.67%)
"Books"	4 (1.0%)	139 (10.98%)
"Music"	143 (35.57%)	385 (30.41%)
"Theatre"	23 (5.72%)	310 (24.49%)
Other	82 (20.4%)	none

Table 3 Frequency of tag classified into classes and events frequencies with respect to classes

Class	Tag Frequency	MAE
"Appointment"	13.18%	0.25
"Art"	16.17%	0.20
"Cinema"	7.96%	0.12
"Books"	1.0%	0.08
"Music"	35.57%	0.3
"Theatre"	5.72%	0.19

Table 4 Tag frequency and MAE values distributed for classes

if users were tagging events related to the appointment class. Moreover, appointments can also be related to concepts that are not modeled in our taxonomy, such as technology and informatics or volunteer organizations. Due to the heterogeneity of this class, we decided that, in the next release of the system, we will consider only the action of tagging as relevant user feedback for this class, and will relate the meaning of the tags to the other classes/subclasses of the taxonomy in the future redesign of the tagging component. Finally, as reported below, most tags were discarded since they were not directly linked to the domain ontology.
We first analyzed 92 user models containing in all 578 tags (210 other tags belonged to user models that were never updated and thus they were discarded). Of these 578 tags, the system dropped 176 (30.4%) since they were not directly related to the classes/subclasses of the domain ontology enriched with MultiWordNet or were invented/unknown.

We should note that in iCITY tags can be either voluntarily inserted by users or suggested by the system. Among the 402 tags we analyzed, 84% were proposed by the system and just clicked on by the users for insertion when tagging the events, while the remaining 16% were inserted by users as free text. Even if most tags are system-generated, we have to specify that the system is not necessarily able to classify correctly the tag in the ontology domain. Indeed tags can belong to the controlled vocabulary, since they are automatically extracted from the event title and linked to the event. In this case the system may not be able to classify the tags in the event ontology. For instance, words of the event titled "U2 concert in turin" are treated as a tag associated with the event, but the system is not able to analyze the meaning of the word "U2", so it automatically discards this tag as unknown in the automatic classification process described above.

The analyzed tags were classified into the event classes displayed in Table 3, see the column Tag Frequency. We have mapped these values on every one of the corresponding 80 users to whom the tag belongs. Thus, we obtained 80 user models reflecting the distribution of the meaning of the tags on the main classes of the event ontology. Values indicating user interests were computed with respect to each class c according to the following formula:

$$interestValue = \frac{|tags_c|}{|allTags|} \tag{8}$$

Class	Tag Frequency	MAE
"Appointment"	9.17%	0.16
"Art"	11.25%	0.29
"Cinema"	5.54%	0.20
"Books"	0.69%	0.14
"Music"	24.74%	0.55
"Theatre"	3.17%	0.16

Table 5 Tag frequency and MAE values distributed for classes, considering users' true ratings for interest values

Class	Tag Frequency	MAE
"Appointment"	15.91%	0.2
"Art"	22.73%	0.2
"Cinema"	2.27%	0.11
"Books"	4.55%	0.07
"Music"	40.91%	0.26
"Theatre"	13.64%	0.19

Table 6 MAE values distributed for classes, considering users' tags inserted as free text

where $tags_c$ are the tags generated by the current user which were mapped to class c, and *allTags* are all the tags generated by the current user.

We then calculated the difference between i) the user interests calculated by the system, reflecting the inferences concerning the user actions, and ii) the user interests reflecting the distribution of the meaning of the tags. To estimate the difference numerically, we calculated MAE. In this case, MAE measures the average absolute deviation between a predicted rating and the distribution of the meaning of the tags for that value. The obtained value was a medium MAE of 0.19, thus not very satisfying. Notice that, again, the scores assigned to each class of interest (*interestValue*) were normalized to sum up to 1. Therefore the value of MAE refers to these normalized values.

We were also curious to analyze the predictions based on many tags and those based on a few, and which classes have low values of MAE. Table 4 thus presents the tag frequency for classes and the corresponding MAE values distributed for classes. As the reader can see, having more tags in a class does not ensure good values of MAE. In fact, the presence of more tags can be due simply to more events belonging to that class and not to more user interest. Thus, we have also calculated the event distribution between classes (see column Event Frequency in Table 3). The comparison between Table 3, Table 4, and 5 shows that the presence of tags in the user model may in part depend on the quantity of events belonging to a given class. For instance, most events belong to the Music class and consequently so do most of the tags. The impact of the tags classified in a given class of the user model therefore needs to be balanced with the number of events belonging to that class with respect to the total number of events.

Afterwards, when available, we also calculated the difference between distribution of the meaning of the tags in the main classes of interests and the users' true ratings for those values. Data for 14 users were available. The obtained value was a medium MAE of 0.25. The average values therefore become slightly worse, but are related to only a limited number of users. Concerning the distribution of these metrics on the event classes, see Table 5.

Finally, we also analyzed the tags inserted by users as free text, for which we obtained a medium MAE of 0.17, quite similar to that which considers both system and user tags. The corresponding value distribution for event classes is displayed in Table 6, as well as tag

Category	Tag frequency
General tags	348 (54.9%)
Specific tags	125 (19.7%)
Spatial tags	66 (10.4%)
Compound tags	55 (8.7%)
Temporal tags	19 (3.0%)
Synonym tags	11 (1.7%)
Subjective tags	8 (1.3%)

Table 7 Tag frequencies with respect to tag categories

frequency. Even in this case, the high frequency of tags does not ensure good MAE results. However, the reader should also remember that most user tags were discarded since they were invented/unknown.

The obtained values demonstrate that the meaning of the tags only partially reflects the user interest in that meaning. Other data need to be considered to refine the values of the user model. From this analysis, we can make some observations about the role of the tags in the update of an overlay user model:

- *Tags themselves are not sufficient to provide a social adaptive system with complete information about the user*. There is a relevant difference between the values distribution of interests present in the user models and those that refer to the classes to which the tags belong. This difference is even more relevant when the interests explicitly expressed by the users are considered. The reasons could be manifold. For instance, the presence of a certain quantity of tags in a class could be due merely to the presence of more events in that class. See the example of the number of tags present in the Music class in Table 4 and the corresponding event frequency in Table 3. Moreover, users could also be not so sincere with respect to their real interests when they are explicitly declaring it: they could declare a strong interest in arts and theater but actually only tag concerts and music;
- A significant part of the tags (20.4%) considered in the classification are not directly related to any concept in the domain ontology, see Table 3, row "Other". Thus, even if a user likes an item, he or she might not necessarily label it using a word related to the domain concept to which the item belongs;
- A significant part of tags may not be classified in any way since users often use invented/unknown tags.

From the above observation we can draw some findings. First, the impact of tags on the user model has to be balanced with other implicit and explicit feedback from the users, even if we still consider them an important indicator of user interests. Second, in order to achieve a complete view of the meaning of the tags, a tag-based user modeling component that classify tags into the domain ontology should extend the overlay user model to other general classes, somehow related to the ones in the domain, or should analyze tags semantically in more detail, using some disambiguation strategies.

Finally, we also analyzed the meaning of tags with respect to the categories described in Section 4. We considered all the tags inserted into the system during the experimentation, namely 788 tags. Of these, we discarded 156 tags (19%), since they were either unknown or foreign words. The results for the remaining 632 tags are shown in Table 7.

These results are quite similar to those presented in Section 4. As suggested by the previous evaluation, we did not classify tags corresponding to subclasses of our ontology as specific tags, but as general tags. Indeed the percentage of general tags increased. With respect

to the previous classification we added: i) the category "*temporal tags*", which groups tags related to a temporal dimension of the events, such as the year/month/day of the event. Note that, with respect to the classification reported in our preliminary evaluation (see Section 4), contextual tags have been split into spatial and temporal tags; ii) the category "compound tags" since a relevant part of the user tags were compound words, such as "art&nature", "orchestralmusic". Classifying tags by part-of-speech, 98% of tags were nouns, while 1% were verbs and another 1% adjectives.

5.4 The final users' satisfaction and their opinions about the system

In this section, we describe how we evaluated users' satisfaction with the system and their opinions about it (goal 3). Explicit user opinions were collected by means of a questionnaire, with the aim of obtaining useful insights into overall user satisfaction and recommendation accuracy. Moreover, users were also questioned about specific HCI aspects.

The questionnaire was distributed to users starting at the end of October 2008 for a period of 15 days.All the 313 registered users were invited by email to respond with their opinions and suggestions for improving iCITY. In order to fill in the questionnaire, users were required to log in to the hosting page using the log-in credentials they used for iCITY. The questionnaire could be filled in by every user only once. Of the 313 registered users, 75 (23.96%) completed the questionnaire.

The questionnaire was partitioned into four sections: the objective of the *first section* was to collect some socio-demographic information from users (age and education), the *second section* dealt with user familiarity with Internet technology, and the *third section* investigated the *user perceptions of the system* and *user interests*. Finally, the *fourth section* was devoted to collecting users' free-text comments about possible improvements for future releases of the system and integration with other social systems.

Questionnaire data were analyzed both quantitatively (questionnaire sections 1-3) and qualitatively (questionnaire section 4). The latter aspect was inspired by the Grounded Theory approach (see Section 2), in order to integrate and reinforce correlational analysis, in accordance with our inductive approach.

5.4.1 The Quantitative Analysis of questionnaires

User demographics. Of the 75 users, 38 were male (50.67%) and 37 female (49.33%), with an age distribution ranging from 15 to 54 years-old. The detailed age distribution was: 10% 15-24, 36% 25-34, 21% 35-44, 32% 45-54.
The level of education was high: 55% graduate, 37% high school, 8% junior high school.
Familiarity with Internet technology. All 75 users were heavy Internet users: 75% of them declared they always use the Internet, while 25% declared they connect often. Concerning the reason for Internet usage, 8% stated mainly for work, 10% mainly during their free time and 82% for both purposes.
Finally, 90% of the users declared they use at least one social software (multiple answers were allowed). More specifically, 24% use del.icio.us, 8% Digg, 52% Flickr, 1% Friendster, 21% LastFm, 35% LinkedIn, 28% MySpace, 7% Pandora, 80% Youtube and 44% Others (Facebook, eBay, Twitter, Listal, Netlog, Elgg, Xing, Slide Share, Yahoo Answer, Yahoo Messenger).
User perceptions of the system. Several questions allowed us to *investigate user satisfaction* with iCITY (goal 4).

Users expressed a positive *overall rating* of their experience with iCITY. The possible scores ranged in a Likert scale from 1 to 5, where 1 indicates an unsatisfactory rating, while 5 indicates an excellent rating. 43% assigned a score of 3, 21% assigned 4, 13% assigned 2, 12% assigned 5, 6% assigned 1, 4% 0. The medium value is 3.07; the standard deviation is 0.25.
A group of questions concerned the *communication of adaptation*. In particular, we asked users about the meaning of icons used to annotate recommended events. 71% declared that the meaning of the colored thumbs-up icons used to visualize the system's prediction about how interesting an event is to a certain user was clear, while the remaining 29% declared that they had not understood their meaning. Concerning the difference between the thumbs-up icons and the stars used to express the rating of the users for every event, 67% stated that it was clear, while the remaining 33% declared it was not intuitive.
As far as *user generated content* is concerned, almost two thirds of users (72%) declared that they had actually noticed that some of the events were posted by other users; among them, 45% considered these events more interesting than those provided via RSS by the official source and 54% found no difference. However, 28% of users apparently did not notice the presence of user-generated content at all.
As regards the *open user model*, users were first asked if they had ever visited the page displaying their user model: 80% answered negatively and only the remaining 20% positively. Of the users who actually viewed their user model, more than a half (53%) stated that they had no difficulty in interpreting it, while 47% reported that it was not clear enough.

5.4.2 The Qualitative Analysis of Questionnaires

All the users' free-text comments were analyzed in accordance with the main guidelines provided by Grounded Theory. As seen in the Related Work section, Grounded Theory is employed to define a theory using an inductive process of collection and analysis of qualitative data [Strauss and Corbin, 1998]. According to the Grounded Theory methodology, collected data may be qualitative or quantitative or a combination of both types, since an interplay between qualitative and quantitative methods is advocated.
More specifically, we based our qualitative analysis on the first two stages involved in the Grounded Theory methodology: i) a constant comparative analysis, the *open coding*, that is an analytical process for identifying conceptual categories and their properties from the analysis of the collected material; ii) a theoretical sampling, the *axial coding*, by which the conceptual categories are related to their subcategories and enriched through coding and integration. We closely examined free text data and compared them for similarities and differences, and we started to accumulate concepts. At the same time, we started the inductive process of the investigation and definition of main categories, subcategories, and variables involved in the phenomenon being studied. After that, the main categories identified in the study and their properties were:

- *Functionalities:* users urge us either to refine or correct some of the pre-existing functionalities and suggest the implementation of completely new ones. As regards adjustments, users asked for: richer information about events and registered users; complete accessibility for visually-impaired people; and detailed and contextual help sections. On the other hand, some of the proposed new functionalities concern, for instance, the possibility to post videos, a search facility which allows the user to retrieve events based on their date, and interoperability with other social networks. Moreover, users would like to be sent emails which remind them of the upcoming events that they added to their favourites.

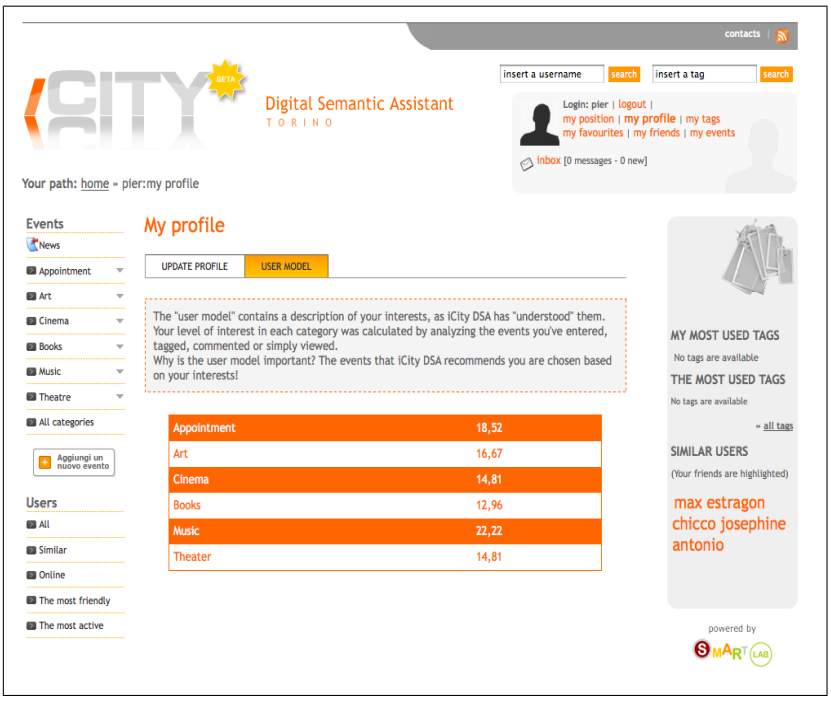

Fig. 6 The open user model

- *Interface adaptivity:* users stressed the need for improvements in the user interface in order to enjoy the personalization aspects fully; in particular, self-explanatory, well characterized and easily distinguishable icons were requested for adaptive annotations (i.e., 33% of users did not understand the difference between stars and thumb-up icons), as well as a clearer, easy-to-understand and more eye-catching visualization for the open user model (see Figure 6).
- *Peer production:* users demand a stronger focus on the aspects of user participation in content creation (peer production), in particular, giving more visibility to those events and contents created by users themselves, in a Web 2.0 perspective.
- *Integration:* users call for the integration of iCITY with the other Web 2.0 systems of the Municipality of Turin in order to obtain more accurate and extensive information about the events occurring in the city. More specifically, they suggested making the systems available on a unique platform, with their services accessible through shared authentication credentials. Moreover, users propose some scenarios of integrated functions, for instance, the possibility of joining iCITY with a system named MappaTO[12], which provides users with the opportunity to create personalized routes in the city of Turin. According to several users, the joining of these two systems would allow users to create personalized maps of the events proposed in iCITY.

The general assessment of the system was quite positive. However, the communication of recommendations was not very successful, and has to be improved in any future re-design. From the qualitative analysis of the results, some very interesting suggestions emerged for

[12] http://www.comune.torino.it/mappato

the future improvement of the system in a socially oriented way, which will be discussed in Section 7. The results of the post-questionnaire analysis were quite interesting also in terms of users' opinions of open user models. In fact, the questionnaires showed that only a few users actually accessed their open user model page and, of these almost half was not very satisfied with its presentation. Given the importance of open user models in social adaptive applications, it is very important to develop strategies to stimulate users to access their open user model page, and to adopt suitable and effective visualizations. In Section 7 we will also present some insights from a study where we expressly investigated these issues.

6 Discussion

General insights. On the whole, the evaluation of iCITY has revealed some interesting results, which deserve discussion.
First of all, it should be noted that out of 2,989 absolute unique visitors only 313 (10.47%) decided to register for the system. This could be explained partially by the fact that iCITY was promoted as an experimental system. Many visitors probably thought it was not worth wasting time on registration for a system with an end date. However, another reason could be that users did not perceive the added value of registration, and limited themselves to browsing cultural events in the non-adaptive version of the system. This is a typical problem related to systems that offer a set of services without registration and advanced services that require registration, as is frequently found in adaptive web systems where a non-adaptive version is often available. Therefore, the suggestions that resulted from the analysis are to make the added value of the registration more obvious and to increase this added value, for example by adding some social features.

At this point, clarification is needed. The social features of iCITY can be split into User-Generated Content features and Social Network features. The former are related to users being able to generate contents; the latter are more concerned with socialization and community aspects. The analysis of the open answers of the questionnaires showed that iCITY users demand a stronger focus on the aspects of both user participation in content creation and social networking, which would probably be a way to increase the advantages of becoming a registered user, and therefore stimulate the number of registrations.

With the objective of deriving general conclusions about the interaction codes of iCITY users, which could be of help to us when planning actions for a future re-design of the system, we tried to define general user profiles. We therefore carefully considered the results of the analysis of the questionnaire, the general observations about user participation, and the findings that emerged from the positive and negative correlations between adaptive web actions and social actions. We summarize here the most interesting interaction codes that emerged:

- Cluster analysis allowed us to identify two interesting types of users, "heavy users" and "email users".
- The "heavy users" cluster is characterized by very different kinds of actions, some related to content visualization (e.g., show_event or map_event_click), some to a few content insertion actions (i.e., the so-called *micro-contributions*, such as add_rating) and some to social networking actions (e.g., add_friend), some of which may imply the insertion of elaborate content (e.g., update_profile). Comparing these observations with the negative correlation we found between single-click actions, such as rating an event, and elaborate text insertion actions, we assumed that the "heavy users" cluster could be

further segmented, distinguishing users who are very active in consuming contents from users who are very active in social networking and generating related content.
- Where content insertion is concerned, we noted that elaborate content is posted only by a minority of users, unless it is aimed at self-presentation and, possibly, networking with other users, as is the case when users update their profile.
- Concerning recommendations, our results show that there is no strong correlation between social and system recommendations. This result suggests that users who appreciate social recommendations are not very interested in system recommendations, and vice versa.
- As regards actions related to the same events, we found that visualizing a recommended event leads users to add it to their favourites (by bookmarking it), and rate, or tag it. We also found that micro-contributions and single-click actions can co-occur on the same event; in particular, users who add an event to their favourites also seem interested in annotating it further through ratings or tags.

Starting from these results, we observed that iCITY users can be described through the following profiles:

- *Consumers*: This kind of user partially corresponds to the "heavy user" who neither performs social networking actions, nor generates elaborate content. Consumers really like social and adaptive features. In particular they follow and like system recommendations, they add and click tags, they rate and add favourites, but they tend not to generate elaborate contents, that is, they add a lot of tags, and micro-contributions but they do not update events. Recommendations are appreciated probably because they represent a shortcut to the fruition of interesting content, which they can quickly annotate with tags or ratings. These users probably like social web sites that do not ask for elaborate content but only for the insertion of tags.
- *Friends for friends' sake*: They partially match "heavy users" who perform mainly social networking actions. Friends for friends' sake are mainly interested in connecting with users, rather than in content generated by the system. They like all the social networking features of the system, such as adding friends, updating their profile and sending messages. At the moment, these users are only partially supported by the system, since iCITY does not offer fully-fledged social networking functions. They would probably appreciate facilities for sharing content with their friends (e.g., recommending an event to a friend) or for interacting with them in real time.
- *Social lurkers*: This kind of user partially corresponds to "email users". Moreover, we considered that visualizing personal messages often leads to visualizing events, and that visualizing socially recommended events often leads to their annotation. Social lurkers really like social actions and resemble "consumers" in that they tend not to generate elaborate content, even if they quite often add tags or favourites. However, they are characterized by their preference for social rather than system recommendations: in fact, these users often check their inbox in order to read the system notifications that update them about the activities of their friends, and then view the events they posted or bookmarked. Moreover, they sometimes visit the profile of other users in order to discover which events they liked. Thus, they seem to be very interested in the activities and preferences of their friends, even if they do not use the system to interact with them directly. They are more interested in users and User Generated Contents than in content provided by an "anonymous" system. To satisfy these users, the adaptivity should give prominence to these social aspects and support users in this activity. At the moment, the iCITY functionalities do not favor this kind of user. Lurkers would probably appreciate

user-based navigation: in addition to a tag cloud, the system could present a "friends cloud", thus making it possible to navigate not only across content but also across users and friends (currently iCITY presents a "similar users cloud"). Moreover, it should also be possible to sort events on the basis of the ratings given by friends.

Since these three profiles were derived from the authors' intuitive understanding of several findings that describe typical behaviors rather than fixed rules, it is quite hard to match each and every user with a single profile. However, according to the partial correspondences that we have outlined among user profiles and user clusters, we can estimate that social lurkers constitute about 30% of users, consumers about 30%, and friends for friend's sake about 15%, while the remaining approximately 25% of users performed only a very few actions in iCITY and could not be attributed to any profile. Moreover, we should also consider that a small percentage of users (around 5%) actively generates more elaborate content, such as the insertion of events, an aspect which is not represented in the proposed user profiles.

In fact, it should be noted that no one profile is characterized by the production of elaborate content, e.g., adding a new event or updating the information about an existing one, while less elaborate social actions seem to characterize all three profiles. Instead, most of the users enjoy the social features of the systems, or appreciate the possibility of exploring the personal pages of their friends. This finding is consistent with Nonnecke et al.'s [2004] claim that content consumers outnumber content producers and with the "90 - 9 - 1" rule formulated by Nielsen[13], which states that the majority of users just consume content produced by others and a small set provides small contributions every now and then, while a very small fraction of users accounts for most of the user-generated content.

It is also interesting to observe that, in other work in the area of social software and online communities, user profiles similar to those we identified are discussed. For example, in Nonnecke et al. [2004] and Preece et al. [2004], the authors refer to "lurkers" and "posters", where lurkers are members of online communities who read, but do not post, and posters are the few members who also post content. With respect to our profiles, both consumers and social lurkers behave as lurkers, but social lurkers prefer to read content which is in some way suggested by other users (e.g., social recommendations). In Prieur et al. [2008], the following user profiles are distinguished with respect to photo-sharing habits in Flickr[14]: "MySpace-like", who use social features, independently of photo sharing, "social media", who share photos and are interested in social interaction around content, and "stockpiling", who share photos, but are not interested in social interaction. While both social media and stockpiling users share some features of posters, who are not represented in our profiles, "MySpace-like" users are similar to friends for friends' sake.

Our findings about social and system recommendations deserve further discussion. If we limit ourselves to comparing their respective values for precision (system: 0.85; social: 0.4) and recall (system: 0.43; social: 0.24), social recommendations show quite a disappointing performance, which is in contrast with the findings of Sinha and Swearingen [2001] and with the flourishing research in the area of social recommendations (see Section 2). However, our results should be weighed in the light of the limitations of the current interface, which does not promote social recommendations and does not highlight user-generated content. If we focus on users who actually visualized social recommendations, on the other hand, we observe that they annotate the recommended event in 44% of cases. Considering that users generate content relatively rarely, this can be considered a good sign: users who had the chance to access social recommendations actually appreciated them. Moreover, it should be

[13] http://www.useit.com/alertbox/participation_inequality.html

[14] http://www.flickr.com/

remembered that our analysis of correlations showed that users who appreciate social recommendations are probably different from users who appreciate system recommendations. In order to support all users, future releases of iCITY should therefore place more emphasis on social recommendations, perhaps providing shortcuts to friend-related content.

Finally, our analysis of tags showed that users exploit different kinds of keywords in order to annotate content of interest. Such tag categories should be taken into account in the design of tag suggestion facilities. For example, general tags should be preferred, since users seem to use them the most frequently. Moreover, support for geo-tagging could be provided, since we also found a significant percentage of spatial tags.

As far as the impact of tags on user modeling is concerned, we found that if only the meaning of tags is considered, accurate inferences on user interests cannot be derived. However, notice that we compared user models based on the meaning of tags alone with user models derived from a large set of actions, although they were analyzed only quantitatively. Thus, we can still claim that tags provide valuable information for user modeling purposes. In order to derive accurate models, however, their impact should be weighed with respect to: i) other actions that provide information about user interests (e.g., ratings, bookmarks, views); ii) factors which could bias their informativeness. As discussed in Section 5.3, the amount of content in each category is likely to influence the number of tags per category, thus influencing tag-based estimations of user interests. A simple solution to deal with this issue would be to normalize the tag-based value indicating user interest in a certain class based on the number of events per class, according to the following Formula:

$$interestValue_{normalized} = \frac{interestValue * classWeight}{totalWeigtht} \tag{9}$$

where *classWeight* is a weight depending on the number of events in a given class c and *totalWeight* is the sum of all class weights. *classWeight* is calculated as $1/|events_c|$. Since reasoning on the meaning of tags provides valuable information for user modeling, further analysis is needed in order to determine how to deal with them correctly.

Insights for the evaluation of social adaptive systems. In the evaluation of the social adaptive system iCITY, we made use of an integrated approach that combines different qualitative and quantitative techniques, in order to perform a comprehensive evaluation of both the social and adaptive aspects of the system. A similar combination of qualitative and quantitative methods was adopted by Fidel and Crandall [1997] with the aim of evaluating users' perceptions of the performance of a filtering system in use at the Boeing Company. Their study comprised three phases. In the first, data about user searching behavior, criteria for identifying relevant items, and user satisfaction with the system were collected by means of observation and interviews. In the second phase, data from the transcribed verbal protocols were analyzed in order to identify criteria for item selection; a questionnaire was developed on this basis. Finally, in the third phase, questionnaire data were statistically analyzed and further interviews were conducted to help interpret the quantitative results.

As far as the evaluation of other social adaptive systems is concerned, in the evaluation described in this paper we specifically focus on two points: i) what is evaluated (evaluation objects), and ii) how it is evaluated (evaluation methods), in order to compare easily iCITY evaluation with the evaluations performed for other systems (see Table 8).

As we have described in detail in this paper, various aspects were taken into account in the evaluation of iCITY, in particular: i) system recommendations quality, ii) social recommendations quality, iii) user satisfaction, iv) user behaviour in the system, and v) accuracy of the tag-based user model.

System	Evaluation Object (what?)	Evaluation Method (how?)
iCITY	i) system recommendations; ii) social recommendations; iii) user satisfaction; iv) user's behavior schemas; v) accuracy of tag-based user model	i) and ii) precision, recall, MAE and RMSE; iii) grounded theory; iv) statistical correlations, sequential analysis, clustering; v) tags classification and MAE and RMSE
Movie Tuner	i) user satisfaction; ii) user interaction with the system	i) statistic analysis; ii) field study of log data
Mypes	i) profile completeness; ii) aggregation benefits; iii) recommendation quality	i) completeness percentage; ii) information gain, entropy, profile overlap; iii) MRR, S@k, P@k
iDYNamicTV	i) recommendation quality; ii) user satisfaction	i) and ii) statistic analysis of survey and log data
SoC-Connect	i) accuracy of inferred user preferences; ii) recommendation quality	i) and ii) qualitative and quantitative analysis of survey and log data
Knowledge Sea II	user satisfaction	statistic analysis of survey data
SUMI	i) scrutability; ii) privacy	i) and ii) controlled experiments with surveys, analysis of survey data

Table 8 Evaluations of social adaptive systems

Data were collected through activity logs and a survey. The quality of the recommendations was evaluated by means of precision, recall, MAE, and RMSE metrics; user answers relating to their satisfaction with the system were qualitatively analyzed, taking inspiration from the Grounded Theory; and user behavior was examined through correlational studies and a sequential analysis. Finally, the accuracy of the tag-based user model was assessed with MAE and RMSE metrics.

To the best of the authors' knowledge, there are very few examples of other evaluations of *social adaptive systems* that use a comparable range of techniques and evaluate a comparable range of objects (see Table 8). An exception is the Movie Tuner [Vig et al., 2011], a novel interface that supports navigation from an item to nearby ones along dimensions represented by tags. In this case, the authors evaluated user satisfaction by means of a statistical analysis of survey data. Moreover, they performed a field study based on activity logs in order to assess user interaction with the system; in terms of this aspect, Movie Tuner evaluation is very similar to that of iCITY.

Conversely, most social adaptive systems perform only an evaluation of some specific aspect. Mypes [Abel et al., 2012] is a user-modeling service that allows the aggregation of public profiles, both tag-based and explicitly defined by users, that are distributed on the Social Web. In this system, the completeness of individual and aggregated user profiles was measured taking into account the profile's percentage of completeness. Information gain, entropy and overlap of the individual profiles were used to assess the benefits of profile aggregation. Final results showed that i) users reveal different types of facets in different systems; the overlap of the individual user profiles across the different systems is rather low; aggregated tag-based user profiles reveal significantly more information about the users than the profiles available in some specific service. Tag and resource recommendation quality were also assessed when Mypes profiles were used in combination with FolkRank, a standard recommendation algorithm for folksonomy systems. Different metrics were exploited: MRR (Mean Reciprocal Rank), which indicates the rank at which the first relevant tag/resource

occurs on average; S@k, which represents the probability that a relevant tag/resource occurs within the first *k* ranks and P@k, which denotes the proportion of relevant items within the first *k* ranks. Final results showed that the consideration of external profile information improves the quality of tag and resource recommendations significantly.

In iDYNamicTV [Carmagnola et al., 2011a], survey data and log data were analyzed with the aim of assessing system recommendation quality, in particular, whether the system correctly infers user interests from user actions, and users' overall satisfaction with the system.

SoC-Connect [Wang et al., 2010] was evaluated with respect to: i) the performance of four machine learning techniques used in the system for learning user preferences on social activities, and ii) the quality of personalized recommendations when different features are used to represent social activities. Data gathered through a questionnaire, as well as social data streams, were analysed qualitatively and quantitatively.

User satisfaction with the system, with particular reference to student opinions about the adopted adaptation techniques, was assessed for Knowledge Sea II [Brusilovsky et al., 2004]. More specifically, student opinions were collected by means of a questionnaire, after they had used the system as well as traditional textbooks during a course at the University of Pittsburgh.

Finally, the evaluation of SUMI [Kyriacou et al., 2009] addressed the scrutability and privacy functionalities of the system. Two different kinds of evaluation were performed, with a series of tasks related to privacy and user-model management through the SUMI website. In the first case, users were guided through a step-by-step process from task to task; in the second, users were free to navigate through the SUMI website. User opinions were collected by means of pre- and post- qualitative questionnaires, including both questions with multiple-choice answers and free-text fields. Furthermore, users were asked to respond to some questions while they were performing controlled tasks.

7 Conclusion

In this paper, we have presented the evaluations we carried out of iCITY, a social, adaptive system that recommends cultural events. In this Section, we summarize our main findings regarding user interactions with a social adaptive recommender system, and, at a meta-level, methodological issues for the evaluation of such systems. We also present a follow-up evaluation that we carried out in order to investigate further some of our findings.

Findings and guidelines on user interactions with a social adaptive system. The first important contributions which emerge from our evaluation are that 1) there is a strong correlation between some social and adaptive features (e.g., clicking on a system recommended event and rating or bookmarking an event and more importantly clicking on a recommended event and adding a tag), 2) there is little correlation between user selection of system and social recommendations (i.e., these two types of recommendations appeal to different users) and 3) the different roles we have observed users can play, namely "consumers", "friends for friend's sake" and "social lurkers" could be taken into account in the design of future social adaptive systems, by integrating features which can support the specific preferences and needs of these different user roles, detailed in Section 6. For example, tagging support suits "consumers", facilities for sharing contents and interacting with friends suit "friends for friend's sake", and a user-based navigation suits "social lurkers". Even if our classification does not consider "posters" [Nonnecke et al., 2004, Preece et al., 2004], we believe that

more active users should also be taken into account in the design of social adaptive system, supporting and rewarding them.

Interesting user requirements for social adaptive applications emerged from the questionnaire. First, users want an improvement in *interface aspects*: more attention to accessibility problems and more visibility of the open user model. Second, they require specific *functionality*: emailing upcoming events and integration with other social networks.
This last point deserves deeper consideration. Web 2.0 users value integration among social systems, since they are used to web sites that aggregate content from various sources. In line with this finding, we suggest that designers of social adaptive systems try to integrate their application with others. Regarding this aspect, iCITY has recently integrated an interoperability module [Cena et al., 2008] that allows it to export tags to another social adaptive application, CHIP [Wang et al., 2008], a system that suggests artworks and virtual tours of the Rijksmuseum of Amsterdam. This is an example of re-use of user interaction data (tags) generated by one application into another one in a similar domain for solving the cold-start problem and providing cross-systems recommendations [Carmagnola et al., 2011b].

The interoperability module maps iCITY tags to the concepts in CHIP, using several standard shared vocabularies, such as Simple Knowledge Organization System (SKOS)[15], Getty AAT, ULAN and general purpose lexical data (WordNet[16]). Other social adaptive systems might allow their users to specify their account on social websites such as Flickr or Delicious[17], import their tags, and map them to domain concepts by means of shared ontologies or vocabularies.
From our experience, we can claim that, to achieve the aim of making interoperability between any social systems possible, a shift towards an extended use of standard ontologies and vocabularies for representing domain content, shared among applications according to a Semantic Web vision, is necessary. In a sense, we could claim that a Semantic Web is needed to support the complete growth of Web 2.0.

Findings and guidelines on social and system recommendations. A valuable insight emerges from our evaluation: users who appreciate traditional system recommendations and users who appreciate social recommendations are often different people. Social adaptive systems whose aim is to target a wide public should offer both. Moreover, we found that, in iCITY, social recommendations performed worse than traditional ones and we explained this result as stemming from interface limitations. Thus, we suggest that, depending on the user profiles detailed above, social adaptive systems should present social recommendations prominently on the home page, or on a dedicated "recommendation" page, as is often the case in traditional recommendations, as well as offering easy and quick ways to access contents liked by friends and other users. If users are offered no support when browsing social recommendations, potentially interesting suggestions are likely to be missed.

Findings and guidelines on the contribution of tags in the user modeling process and on tag usage. Regarding the contribution of tags in the user modeling process, our results show that the meaning of the tags only partially reflects user interests for the corresponding class in a domain taxonomy. On the one hand, this result suggests that various actions should be considered, rather than tags alone, as indicators of user interests in social adaptive systems. On the other hand, it demands further analysis aimed at investigating possible biases to tag effectiveness: as suggested above, the impact of the number of events per category has to be considered. In addition, we were not able to profit from tags which did not directly map

[15] http://www.w3.org/2006/07/SWD/wiki/SkosDesign/ConceptualMapping/ProposalOne
[16] http://wordnet.princeton.edu/
[17] http://www.delicious.com/

to concepts represented in the domain taxonomy. Extended use of a lexical database might help to disambiguate such tags.

Regarding our findings about tag usage, our results are quite consistent with those of our previous analysis, suggesting that the classification we propose is stable and can explain tagging behavior well. If tags are to be recommended, it is advisable to take into account how different types of tag are used. For example, we observed that users, for the most part, adopt general terms. We also found a relatively large number of spatial and temporal tags, which, however, seem to be strongly related to the domain of tagged contents, i.e., cultural events. Tag recommendations that suggest popular types of tags are more likely to be appreciated than tag recommendations suggesting less common keywords, such as specific terms which might be useful to only a minority of users. Further analysis could concentrate on mapping different types of tag to different user profiles, in order to allow more fine-grained recommendations.

Concerning the impact of tags on the user model, and in particular the meaning of tags, we conclude that tags themselves are not sufficient to provide a social adaptive system with complete information about the user model. They have to be analyzed, giving them light emphasis relative to other actions that indicate user interests. However, users who follow recommendations insert many tags, and thus this correlation suggests that, in a social adaptive web site, tags are an important user action. They cannot, therefore, be ignored.

Guidelines for the evaluation of social adaptive web systems. From our experience we can derive some guidelines for the evaluation of such systems. We can state that indirect observational methods can provide two contributions: they allow one to study and analyze the behavior of the users on a web site; and they allow one to improve the results of evaluations that have more limited and specific goals, such as measuring the accuracy of recommendations, using well-established metrics such as precision, recall, MAE, and so on. Many variables can influence these results and a multifaceted evaluation can reveal them. It is important to note that all the possible variables that influence user behavior need to be taken into account. In addition, inductive methods of analysis should be useful in terms of letting new phenomena emerge in the absence of preconceptions, and of proposing new general understandings of reality. Indirect observational methods, as well as a quantitative analysis of the collected user opinions, could be of use in the final evaluation phase. However, users do not always tell the truth, and opinions should therefore always be confirmed by real actions.

Suggestions for future releases of iCITY. We are planning to redesign iCITY, based on our general findings about user interactions with a social adaptive recommender system. New functionalities will be introduced to satisfy the explicitly expressed user needs, as well as to support the three user roles we have identified. In particular, the ideas we want to introduce in the future releases of iCITY are:

- Functions which should satisfy explicitly expressed user needs:
  - Sending email reminders for upcoming bookmarked events;
  - Integrating iCITY with other social networks and with other social systems that support tourists and citizens in the municipality of Turin.
- Support for specific user roles:
  - *Consumers*: Offering an improved tagging facility that suggests tags related to general concepts, time, and space;
  - *Friends for friends'sake*: Allowing users to suggest events to one or more friends; allowing them to exchange real time messages with online friends;

- *Social lurkers*: Improving user-based navigation by including a friend cloud in addition to the current similar user cloud. Other user clouds might be considered, for example, one presenting users who are very active (e.g., generate a lot of content). Content recommendations should consider the friend-generated content and the content appreciated by friends;
- *Posters*: Offering improved facilities for content insertion and some form of reward inside the community.

Moreover, we plan to redesign the iCITY homepage so that lists of both system and social recommendations are displayed prominently.

Some of the findings that emerged from the evaluation of iCITY cannot be translated immediately into specific ideas for redesign and seem to require some further investigation instead.

First, one third of users stated that they could not understand the semantic difference between the two types of icon (thumbs-up and stars) that we used to annotate recommended events; a careful revision of the visual cues used to provide adaptive annotations is therefore required.

Second, results from the post-usage questionnaire analysis showed that almost half the users who accessed their open user model (26%) were not satisfied with the way it was externalized in our system. This negative result prompted us to plan a redesign of the iCITY open user model, and to carry out a follow-up evaluation that specifically focuses on this topic.

Follow-up evaluation. Open user models play a fundamental role in increasing users' acceptance of social adaptive applications. With the objective of identifying specific suggestions for future releases of iCITY, as well as of deriving general guidelines, we carried out a pilot experiment that investigated user preferences for widgets that could be used for visualizing open user models.

We chose three widgets that differ in both their granularity and visual appearance: a 3-point thumb rating scale (human metaphor), a 5-point star rating scale (cultural metaphor, both playful and professional connotations) and a 10-point slider rating scale (technological metaphor). We hypothesized that user preferences for widgets change on the basis of: 1) user personality traits and 2) the topics that the users are evaluating. We designed a within-subjects, multiple factor (personality traits, topic) experiment. We selected 32 subjects, 15-54 years-old, among colleagues and students at the Computer Science Department following an availability sampling strategy. Users were asked to choose their preferred widgets for *modifying an open user model*, with respect to: i) their preferences for a series of categories of events corresponding to the classes of the domain taxonomy; ii) the subcategories of such categories and their trust in them; iii) a list of related users (i.e., friends and similar users) in iCITY.

The most relevant finding of our investigation is that user choices for widgets depend on the topic the user has to evaluate. This may be due to the fact that some topics are social, that is, they refer to the social network of the user, and others are not, that is, they refer to a user's personal interests. We found a significant correlation between the social value of the topic and the preferred widget ($X^2(2) = 20,595$; $p < 0,001$). Most users (53,13%) chose the stars (mode), followed by the sliders (37,5%) for evaluating their preference for domain categories and subcategories (non-social topics), while they chose the thumbs (48,9%, mode), followed by the stars (41,9%) for evaluating their trust in similar users and friends (social topic). To generalize, users seem to prefer widgets based on a scale with a coarse or medium granularity, and having a human-like or playful visual appearance, for topics of social value. On the other hand, they seem to prefer widgets based on a scale with a medium

or very fine granularity, and having a professional or technological visual appearance, when the evaluated topic has no social value.

This finding allows us to define another *guideline for developers of social adaptive systems* who need to *externalize user models*. Users could be provided with different widgets based on the topic they are evaluating. In particular, according to our results, different widgets should be offered for evaluating social topics and for evaluating non-social topics. For example, in a social adaptive system in the domain of recipes, a thumb rating scale might be provided for evaluating recipes posted by users, while a 5-point or 10-point star rating scale might be provided for evaluating recipes posted by the editorial staff, that is, by the "system".

As for the iCITY open user model, we are planning to design an improved version where users are offered a 5-point star rating scale for evaluating event categories and subcategories, and a 3-point thumb rating scale for evaluating their trust in friends and similar users.

References

F. Abel, E. Herder, G.-J. Houben, N. Henze, and D. Krause. Cross-system user modeling and personalization on the social web. *This issue*, 2012.

R. Bakeman and J. M. Gottman. *Observing Interaction: An Introduction to Sequential Analysis*. Cambridge University Press, 1997.

T. Barker, S. Jones, C. Britton, and D. Messer. The use of a co-operative student model of learner characteristics to configure a multimedia application. *User Model. User-Adapt. Interact.*, 12(2-3):207–241, 2002.

P. Brusilovsky. Methods and techniques of adaptive hypermedia. *User Model. User-Adapt. Interact.*, 6(2-3):87–129, 1996.

P. Brusilovsky, P. Karagiannidis, and D. Sampson. The benefits of layered evaluation of adaptive applications and services. In D. N. C. S. Weibelzahl, editor, *Empirical Evaluation of Adaptive Systems*, volume Proc. of Workshop At the Eighth International Conference on User Modeling, UM2001, pages 1–8, 2001.

P. Brusilovsky, G. Chavan, and R. Farzan. Social adaptive navigation support for open corpus electronic textbooks. In P. De Bra and W. Nejdl, editors, *Proc. of Adaptive Hypermedia and Adaptive Web-Based Systems, AH04*, volume 3137 of *Lecture Notes in Computer Science*, pages 176–189. Springer Berlin / Heidelberg, 2004.

F. Carmagnola, F. Cena, L. Console, O. Cortassa, C. Gena, A. Goy, I. Torre, A. Toso, and F. Vernero. Tag-based user modeling for social multi-device adaptive guides. *User Modeling and User-Adapted Interaction*, 18(5):497–538, 2008.

F. Carmagnola, F. Cena, L. Console, P. Grillo, M. Perrero, R. Simeoni, and F. Vernero. Supporting content discovery and organization in networks of contents and users. *Multimedia Systems*, 17:199–218, 2011a. ISSN 0942-4962. URL http://dx.doi.org/10.1007/s00530-010-0219-4. 10.1007/s00530-010-0219-4.

F. Carmagnola, F. Cena, and C. Gena. User model interoperability: a survey. *User Modeling and User-Adapted Interaction*, 21:285–331, 2011b. ISSN 0924-1868. URL http://dx.doi.org/10.1007/s11257-011-9097-5. 10.1007/s11257-011-9097-5.

F. Cena, C. Gena, and S. Modeo. How to communicate recommendations? Evaluation of an adaptive annotation technique. In M. F. Costabile and F. Paternò, editors, *INTERACT*, volume 3585 of *Lecture Notes in Computer Science*, pages 1030–1033. Springer, 2005. ISBN 3-540-28943-7.

F. Cena, F. Carmagnola, O. Cortassa, C. Gena, Y. Wang, N. Stash, and L. Aroyo. Tag interoperability in cultural web-based applications. In P. Brusilovsky and H. C. Davis, editors, *Hypertext*, pages 221–222. ACM, 2008. ISBN 978-1-59593-985-2.

D. N. Chin. Empirical evaluation of user models and user-adapted systems. *User Modeling and User-Adapted Interaction*, 11(1-2):181–194, 2001. ISSN 0924-1868.

R. Damiano, C. Gena, V. Lombardo, F. Nunnari, and A. Pizzo. A stroll with carletto: adaptation in drama-based tours with virtual characters. *User Modeling and User-Adapted Interaction*, 18(5):417–453, 2008.

M. de Gemmis, P. Lops, G. Semeraro, and P. Basile. Integrating tags in a semantic content-based recommender. In *Proceedings of the 2008 ACM conference on Recommender systems*, RecSys '08, pages 163–170, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-093-7. doi: http://doi.acm.org/10.1145/1454008.1454036. URL http://doi.acm.org/10.1145/1454008.1454036.

J. Diederich and T. Iofciu. Finding Communities of Practice from User Profiles Based On Folksonomies. In *Proceedings of the 1st International Workshop on Building Technology Enhanced Learning solutions for Communities of Practice (TEL-CoPs'06)*, Oct. 2006.

R. Fidel and M. Crandall. Users' perception of the performance of a filtering system. In *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '97, pages 198–205, New York, NY, USA, 1997. ACM. ISBN 0-89791-836-3. doi: http://doi.acm.org/10.1145/258525.258568. URL http://doi.acm.org/10.1145/258525.258568.

J. Freyne, S. Berkovsky, E. M. Daly, and W. Geyer. Social networking feeds: recommending items of interest. In *Proc. of the fourth ACM conference on Recommender systems, RecSys'10*, pages 277–280, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-906-0. doi: http://doi.acm.org/10.1145/1864708.1864766. URL http://doi.acm.org/10.1145/1864708.1864766.

C. Gena. Methods and techniques for the evaluation of user-adaptive systems. *Knowledge Engineering Review*, 20(1):1–37, 2005.

C. Gena and S. Weibelzahl. Usability engineering for the adaptive web. In P. Brusilovsky, A. Kobsa, and W. Nejdl, editors, *The Adaptive Web*, volume 4321 of *Lecture Notes in Computer Science*, pages 720–762. Springer, 2007. ISBN 978-3-540-72078-2.

N. Good, J. B. Schafer, J. A. Konstan, A. Borchers, B. M. Sarwar, J. L. Herlocker, and J. Riedl. Combining collaborative filtering with personal agents for better recommendations. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and Eleventh Conference on Innovative Applications of Artificial Intelligence, AAAI/IAAI 1999*, pages 439–446, Orlando, Florida, July 18-22 1999. AAAI Press / The MIT Press.

I. Guy, N. Zwerdling, D. Carmel, I. Ronen, E. Uziel, S. Yogev, and S. Ofek-Koifman. Personalized recommendation of social software items based on social relations. In *RecSys '09: Proceedings of the third ACM conference on Recommender systems*, pages 53–60, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-435-5. doi: http://doi.acm.org/10.1145/1639714.1639725.

J. L. Herlocker, J. A. Konstan, L. Terveen, and J. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transaction Information System*, 22(1):5–53, 2004. ISSN 1046-8188. doi: http://doi.acm.org/10.1145/963770.963772.

K. Höök. Evaluating the utility and usability of an adaptive hypermedia system. In *Proc. International Conference on Intelligent User Interfaces, IUI97*, pages 179–186. ACM Press, 1997. doi: http://doi.acm.org/10.1145/238218.238320.

A. Jameson. Adaptive interfaces and agents. In *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*, pages 305–330,

Hillsdale, NJ, USA, 2003. L. Erlbaum Associates Inc. ISBN 0-8058-3838-4.

C. Karagiannidis and D. G. Sampson. Layered evaluation of adaptive applications and services. In *Proc. of the International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, AH00*, pages 343–346, London, UK, 2000. Springer-Verlag. ISBN 3-540-67910-3.

J. Kay. Scrutable adaptation: Because we can and must. In V. P. Wade, H. Ashman, and B. Smyth, editors, *AH*, volume 4018 of *Lecture Notes in Computer Science*, pages 11–19. Springer, 2006. ISBN 3-540-34696-1.

G. Keppel, W. H. Saufley, and H. Tokunaga. *Introduction to Design and Analysis: a Student's Handbook*. W H Freeman and Co., 1998.

H.-N. Kim and A. El Saddik. Exploring social tagging for personalized community recommendations. *This issue*, 2012.

A. Kobsa, J. Koenemann, and W. Pohl. Personalized hypermedia presentation techniques for improving online customer relationship. *The Knowledge Engineering Review*, 16(2): 111–115, 2001.

D. Kyriacou, H. C. Davis, and T. Tiropanis. A (multi'domain'sional) scrutable user modelling infrastructure for enriching lifelong user modelling. In *Proceedings of Lifelong User Modelling Workshop AT conference UMAP 2009, Trento*, pages 46–54, 2009.

S. K. Loizou and V. Dimitrova. Adaptive notifications to support knowledge sharing in close-knit virtual communities. *This issue*, 2012.

C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *17th Conference on Hypertext and Hypermedia HYPERTEXT '06*, pages 31–40. ACM, 2006.

M. R. McLaughlin and J. L. Herlocker. A collaborative filtering algorithm and evaluation metric that accurately model the user experience. In M. Sanderson, K. Jrvelin, J. Allan, and P. Bruza, editors, *SIGIR 2004: Proc. of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004*, pages 329–336. ACM, 2004.

E. Michlmayr and S. Cayzer. Learning user profiles from tagging data and leveraging them for personal(ized) information access. In *Proceedings of the Workshop on Tagging and Metadata for Social Information Organization, 16th International World Wide Web Conference*, 2007.

R. Y. Nakamoto, S. Nakajima, J. Miyazaki, S. Uemura, H. Kato, and Y. Inagaki. Reasonable tag-based collaborative filtering for social tagging systems. In *Proceedings of the 2nd ACM workshop on Information credibility on the web*, WICOW '08, pages 11–18, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-259-7. doi: 10.1145/1458527.1458533. URL `http://doi.acm.org/10.1145/1458527.1458533`.

A. Nauerz, F. Bakalov, M. Welsch, and B. Konig-Ries. New tagging paradigms for content recommendation in web 2.0 portals. In *Proceedings of International Workshop on Adaptation and Personalization for Web 2.0 (AP-WEB 2.0 2009)*, pages 143–147, 2009.

J. Nielsen. Risks of quantitative studies. http://www.useit.com/alertbox/20040301.html, Accessed on Jan 2009, March 2004.

W. T. Niu and J. Kay. Pervasive personalisation of location information: Personalised context ontology. In *Proceedings of the 5th international conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, AH '08, pages 143–152, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-70984-8.

B. Nonnecke, J. Preece, and D. Andrews. What lurkers and posters think of each other. In *Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 7 - Volume 7*, HICSS '04, pages 70195.1–,

Washington, DC, USA, 2004. IEEE Computer Society. ISBN 0-7695-2056-1. URL http://dl.acm.org/citation.cfm?id=962755.963094.

A. Paramythis and S. Weibelzahl. A decomposition model for the layered evaluation of interactive adaptive systems. In *User Modeling*, volume 3538 of *Lecture Notes in Computer Science*, pages 438–442. Springer, 2005.

A. Paramythis, P. Totter, and P. Stephanidis. A modular approach to the evaluation of adaptive user interface. In S. Weibelzahl, D. Chin, , and G. Weber, editors, *Proceedings of the First Workshop on Empirical Evaluation of Adaptive Systems*, pages 9–24, Sonthofen, Germany, 2001.

A. Paramythis, S. Weibelzahl, and J. Masthoff. Layered evaluation of interactive adaptive systems: framework and formative methods. *User Modeling and User-Adapted Interaction*, 20:383–453, 2010. ISSN 0924-1868. URL http://dx.doi.org/10.1007/s11257-010-9082-4. 10.1007/s11257-010-9082-4.

P. Pirolli and S. Kairam. A knowledge-tracing model of learning from a social tagging system. *This issue*, 2012.

J. J. Preece, B. Nonnecke, and D. Andrews. The top five reasons for lurking: improving community experiences for everyone. *Computers in Human Behavior*, 20(2):201–223, 2004.

C. Prieur, D. Cardon, J.-S. Beuscart, N. Pissard, and P. Pons. The stength of weak cooperation: A case study on flickr. *CoRR*, abs/0802.2317, 2008.

P. Rizzo, H. Lee, E. Shaw, W. L. Johnson, N. Wang, and R. E. Mayer. A semi-automated wizard of oz interface for modeling tutorial strategies. In *In Proc. of International Conference on User Modeling, UM 2005*, pages 174–178, 2005. doi: 10.1007/11527886. URL http://dx.doi.org/10.1007/11527886.

B. Sarwar, G. Karypis, J. Konstan, and J. Reidl. Item-based collaborative filtering recommendation algorithms. In *Proc. of the 10th international conference on World Wide Web*, pages 285–295. ACM Press New York, NY, USA, 2001.

B. Shapira, L. Rokach, and S. Freilichman. Utilizing facebook single and cross domain data for recommendation systems. *This issue*, 2012.

A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke. Personalized recommendation in social tagging systems using hierarchical clustering. In *Proceedings of the 2008 ACM conference on Recommender systems*, RecSys '08, pages 259–266, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-093-7. doi: http://doi.acm.org/10.1145/1454008.1454048. URL http://doi.acm.org/10.1145/1454008.1454048.

R. R. Sinha and K. Swearingen. Comparing recommendations made by online systems and friends. In *DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries*, 2001.

A. Strauss and J. Corbin. *Basics of Qualitative Research : Techniques and Procedures for Developing Grounded Theory*. SAGE Publications, September 1998. ISBN 0803959400.

M. Szomszor, H. Alani, I. Cantador, K. O'Hara, and N. Shadbolt. Semantic modelling of user interests based on cross-folksonomy analysis. In A. P. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. W. Finin, and K. Thirunarayan, editors, *International Semantic Web Conference*, volume 5318 of *Lecture Notes in Computer Science*, pages 632–648. Springer, 2008. ISBN 978-3-540-88563-4.

Totterdell and Boyle. Adaptive user interfaces. In D. Browne, A. Totterdell, and D. Norman, editors, *The Evaluation of Adaptive Systems*, pages 161–194, London, 1990. Academic Press.

M. van Setten, R. Brussee, H. van Vliet, L. Gazendam, Y. van Houten, and M. Veenstra. On the importance of "Who Tagged What". In *Workshop on the Social Navigation and Community based Adaptation Technologies*, pages 552–561, Dublin, Ireland, 2006. URL http://www.sis.pitt.edu/%7Epaws/SNC_BAT06/crc/vansetten.pdf.

J. Vig, S. Sen, and J. Riedl. Navigating the tag genome. In P. Pu, M. J. Pazzani, E. André, and D. Riecken, editors, *IUI*, pages 93–102. ACM, 2011. ISBN 978-1-4503-0419-1.

Y. Wang, N. Stash, L. Aroyo, P. Gorgels, L. Rutledge, and G. Schreiber. Recommendations based on semantically enriched museum collections. *Journal of Web Semantics*, 6(4): 283–290, 2008.

Y. Wang, J. Zhang, and J. Vassileva. Personalized recommendation of integrated social data across social networking sites. In *Proc. of SASweb workshop, at UMAP 2010*, pages 19–30, 2010.

S. Weibelzahl. Evaluation of adaptive systems. In *User Modeling*, volume 2109 of *Lecture Notes in Computer Science*, pages 292–294. Springer, 2001.

S. Weibelzahl. *Evaluation of Adaptive Systems. Dissertation*. University of Trier, Germany, Trier, 2003.

S. Weibelzahl and C. Lauer. Framework for the evaluation of adaptive CBR-systems. In U. Reimer, S. Schmitt, and I. Vollrath, editors, *Proc. of the 9th German Workshop on Case-Based Reasoning (GWCBR01), Aachen: Shakerm*, pages 254–263, 2001.

S. Weibelzahl and G. Weber. A database of empirical evaluations of adaptive systems. In *Proc. of Workshop Lernen - Lehren - Wissen - Adaptivity (LLWA 01), Research Report in Computer Science Nr. 763, University of Dortmund*, pages 302–306, 2001.

V. Zanardi and L. Capra. Social ranking: uncovering relevant content using tag-based recommender systems. In P. Pu, D. G. Bridge, B. Mobasher, and F. Ricci, editors, *Proc. of International Conference on Recommender Systems, RecSys08*, pages 51–58. ACM, 2008. ISBN 978-1-60558-093-7.

Y. Zhen, W.-J. Li, and D.-Y. Yeung. Tagicofi: tag informed collaborative filtering. In L. D. Bergman, A. Tuzhilin, R. D. Burke, A. Felfernig, and L. Schmidt-Thieme, editors, *RecSys*, pages 69–76. ACM, 2009. ISBN 978-1-60558-435-5.

A. Zimmermann and A. Lorenz. Listen: a user-adaptive audio-augmented museum guide. *User Modeling and User-Adapted Interaction*, 18:389–416, November 2008. ISSN 0924-1868. doi: 10.1007/s11257-008-9049-x. URL http://portal.acm.org/citation.cfm?id=1455853.1455878.

Appendix 1

Correlations significant at 0.01 level*:

- adding a tag and adding a comment (r=1.00, significant at the 0.01 level);
- adding a favourite and clicking on a tag (r=1.00, significant at the 0.01 level);
- adding a favourite and changing the recommendation criteria (r=1.0, significant at the 0.01 level);
- adding a friend and clicking on a tag (r=1.00, significant at the 0.01 level);
- adding a friend and open a message of a friend (r=1.0, significant at the 0.01 level);
- adding an event and localizing a user (r=1.0, significant at the 0.01 level);
- clicking on an event and opening a message (r=1.00, significant at the 0.01 level);
- clicking on a category of navigation and clicking on the map (r=-1.00, significant at the 0.01 level);
- clicking on a event and localizing a user position (r= 0.997, significant at the 0.01 level);
- clicking on a social recommendation and clicking on a tag (r= 0.990 significant at the 0.01 level);
- adding a friend and updating the profile (r=0.988, significant at the 0.01 level);
- clicking on a recommended event and clicking on a tag (r=0.977, significant at the 0.01 level);
- clicking on an event and clicking on a tag (r=0.966, significant at the 0.01 level);
- adding a tag and clicking on a tag (r=0.951, significant at the 0.01 level);
- adding a favourite event and rating an event (r=0.934, significant at the 0.01 level);
- clicking on a recommended event and rating an event (r=0.929, significant at the 0.01 level);
- adding a tag and rating an event (r=0.890, significant at the 0.01 level);
- adding a tag and adding a favourite (r=0.884, significant at the 0.01 level);
- adding a favourite and clicking on a recommended event (r=0.866, significant at the 0.01 level);
- *adding a favourite and clicking on an event (*r=*0.863, significant at the 0.01 level)**;*
- *clicking on an event and rating an event (*r=*0.863, significant at the 0.01 level)**;*
- clicking on an event and clicking on a recommended event (r=0.850, significant at the 0.01 level);
- *adding a tag and clicking on an event (*r=*0.790, significant at the 0.01 level)**;*
- adding a tag and clicking on a recommended event (r= 0.773 significant at the 0.01 level);
- adding a favourite and clicking on a social recommendation (r= 0.750 significant at the 0.01 level);
- adding a tag and clicking on a social recommendation (r= 0.570 significant at the 0.01 level);
- clicking on a recommended event and clicking on a social recommendation (r=0.470, significant at the 0.01 level);
- clicking on an event and clicking on a social recommendation (r=0.440, significant at the 0.01 level);
- adding a favourite event and updating the profile (r=-1.00, significant at the 0.01 level);
- adding a favourite event and updating an event (r=-1.00, significant at the 0.01 level);
- adding an event and clicking on the map (r=-1.00, significant at the 0.01 level);
- adding an event and updating the profile (r=-1.0, significant at the 0.01 level);
- adding a comment and clicking on a subcategory (r=-1.00, significant at the 0.01 level);

- changing recommendation criteria and opening a message (r=-1.00, significant at the 0.01 level);
- changing the recommendation criteria and opening a message (r=-1.0, significant at the 0.01 level);
- clicking on the map and opening a message (r=-1.00, significant at the 0.01 level);
- clicking on a subcategory and clicking on the map (r=-1.0, significant at the 0.01 level);
- rating an event and sending a message (r=-1.00, significant at the 0.01 level);
- rating an event and updating an event (r=-1.00, significant at the 0.01 level);

* Note that for all the correlations 2-tailed tests were performed.
** Meaningless correlations due to interface constraints.

Authors Biographies

(1) Cristina Gena is Assistant Professor at the Department of Computer Science of the University of Torino, working in the area of intelligent user interfaces. She completed her Ph.D. in Communication Science (University of Torino) in 2003, with a thesis on the evaluation of user-adaptive systems.Her current research activities address user modeling, adaptive web systems and their evaluation, context-aware systems, semantic web, web 2.0, usability and interaction design. Her contribution is based on experiences gained within these fields.

(2) Federica Cena is Assistant Professor at the he Department of Computer Science of the University of Torino (Italy). She is working on user modeling, personalization and ubiquitous computing. In the last years, she has been studying the implications of Social Web and Semantic Web for user modeling. She has served as a program committee member on several workshops and conferences. She was program co-chair of the International Workshop on Web 3.0 at Hypertext 2009 and of the Adaptation in Social and Semantic Web Workshop Series at the UMAP conference. She is guest editor of the special issue Social Semantic Adaptive Web on the ACM Transactions on Intelligent Systems and Technology journal, and is a member of the Editorial Board of the international journals "Advances in Internet of Things" and Artificial Intelligence Research".

(3) Fabiana Vernero is a short-term researcher at the Department of Computer Science, University of Torino. She completed her Ph.D. in Computer Science (University of Torino) in 2011. Her main interests lie in the areas of Human-Machine Interaction, Interaction Design and Social Recommender Systems.

(4) Pierluigi Grillo is a Ph.D. Student in Computer Science at the Department of Computer Science of the University of Torino. His research interests are in the area of intelligent user interfaces, mobile computing, ubiquitous computing and user modeling. Currently he is working in the field of Internet of Things.