User Model User-Adap Inter (2007) 17:5–40
DOI 10.1007/s11257-006-9019-0

ORIGINAL PAPER

A hybrid approach for improving predictive accuracy of collaborative filtering algorithms

George Lekakos · George M. Giaglis

Received: 1 November 2005 / Accepted in revised form: 11 July 2006 /
Published online: 25 January 2007

Abstract Recommender systems represent a class of personalized systems that aim at predicting a user's interest on information items available in the application domain, operating upon user-driven ratings on items and/or item features. One of the most widely used recommendation methods is collaborative filtering that exploits the assumption that users who have agreed in the past in their ratings on observed items will eventually agree in the future. Despite the success of recommendation methods and collaborative filtering in particular, in real-world applications they suffer from the insufficient number of available ratings, which significantly affects the accuracy of prediction. In this paper, we propose recommendation approaches that follow the collaborative filtering reasoning and utilize the notion of lifestyle as an effective user characteristic that can group consumers in terms of their behavior as indicated in consumer behavior and marketing theory. Emanating from a basic lifestyle-based recommendation algorithm we incrementally proceed to the development of hybrid recommendation approaches that address certain dimensions of the sparsity problem and empirically evaluate them providing further evidence of their effectiveness.

Keywords Recommender systems · Collaborative filtering · Personalization · Lifestyle

1 Introduction

In many occasions in our everyday life, we become active seekers or passive receivers of information in order to make selections, choices or purchase decisions. However,

G. Lekakos (✉) · G. M. Giaglis
Department of Management Science and Technology,
Athens University of Economics and Business, 47 Evelpidon Str., 11362 Athens, Greece
e-mail: glekakos@aueb.gr

G. M. Giaglis
e-mail: giaglis@aueb.gr

our experiences and knowledge often do not suffice to process and evaluate the vast amount of available information. This information overload problem becomes even greater in the age of easily accessible digital information. It is a common practice to reduce the information processing workload by exploiting the experiences of friends, colleagues, family, or professionals (Resnick and Varian 1997). Malone et al. (1987) observed the "social filtering" process followed by employees in an organizational context to prioritize e-mail messages sent by colleagues who had some form of personal or organizational relationship to them. Goldberg et al. (1992) used the term "collaborative filtering" in Tapestry, an e-mail filtering system, to denote that "people collaborate to help each other perform filtering by recording their reactions to documents they read." Resnick and Varian (1997) coined the term "recommender systems" for systems where "people provide their recommendations which the system then aggregates and directs to appropriate recipients." Nowadays, recommender systems have expanded their scope and the approaches utilized to produce the recommendations and refer to systems that "produce individualized recommendations as output or have the effect of guiding the user in personalized way to interesting or useful objects in a large space of possible options" (Burke 2002).

Recommender systems have proven very useful in several domains such as movies (Alspector et al. 1997; Good et al. 1999), news (Resnick et al. 1994; Maybury 2001), and e-commerce product recommendation applications, including successful commercial systems such as Amazon.com and eBay (Schafer et al. 2001). Recommendation methods operate upon user ratings on observed items and/or item features making predictions concerning users' interest on unobserved items. However, in most cases in particular in real-world applications, the ratio of rated items to the total of available items is very low. The absence of a sufficient amount of available ratings significantly affects recommendation methods reducing the accuracy of prediction. The sparsity of ratings problem is particularly important in domains with large or continuously updated list of items as well as a large number of users. The sparsity problem may occur when either none or few ratings are available for the target user, or for the target item that prediction refers to, or for the entire database in average. Different treatments are required and different prediction techniques must be employed depending on the sparsity conditions, making the selection of an appropriate approach a cumbersome task. Current personalization approaches are limited in the sense that they address specific aspects of the above problem (Herlocker and Konstan 2001; van Setten et al. 2002).

Along these lines we propose recommendation approaches that utilize the notion of lifestyle as a user characteristic that can be effectively exploited to overcome the sparsity problem. Indeed, consumer behavior theory suggests that lifestyle is a significant predictor of consumer's behavior. In adaptive and recommender systems literature, the value of personality traits and in particular lifestyle as personalization feature has been acknowledged (Ardissono and Goy 2000; Brusilovsky 2001) but it has been considered in very few studies (Krulwich 1997; Ardissono et al. 2001). In the present research, the proposed algorithmic implementations manage different aspects of the sparsity problem (in different conditions) and are finally integrated into a single personalization strategy.

The present research is applied on the domain of digital interactive television advertisements. The task of personalizing interactive advertisements is considered as recommending audiovisual information items concerning products, services, or information to individual viewers (users). The advances in digital interactive television

technology (Milenkovic 1998) including set-top boxes with the ability to store and process data, content, and interactive applications enable the development and application of personalization methods and techniques in this context.

The rest of the paper is organized as follows. In the next section, a review of the relevant literature is presented while in Sect. 3, the notion of lifestyle is introduced and existing lifestyle segmentation methods are discussed. In Sect. 4, an initial recommendation approach as a direct implementation of existing lifestyle segmentation methods is presented. In Sect. 5, the basic recommendation algorithm based on lifestyle is presented as well as two more hybrid approaches that extend the above algorithm beyond high-sparsity conditions. In Sect. 6, a meta-learning hybridization technique, which presents item-level sensitivity is proposed. In Sect. 7, the proposed approaches are aggregated into a personalization strategy that operates upon a given set of sparsity conditions and outputs the appropriate recommendation approach, while in the last section of the article conclusions and future research directions are discussed.

2 Recommendation approaches

The recommendation task refers to the prediction of a user's interest for a specific information item (e.g., books, movies, music, products). The user and the item the prediction refers to are indicated as the target user and the target item, respectively. The recommendation process takes as input an expression of users' interest on observed items and/or item features and typically machine learning techniques are applied to make predictions of interest for unobserved items. The personalization effect is then visualized using various techniques, in accordance to the characteristics of the application domain, such as presenting a ranked list of relevant items, recommending the top-*n* relevant items, or providing navigation support by ordering the relevant items (Hollink et al. 2006).

Recommender systems can be classified upon different features, such as the type of data utilized in the recommendation process, the data acquisition mechanism, the output produced, and so on. However, of primary importance are the recommendation approaches, which in combination with domain characteristics (such as the user's goal or task, or the available interaction mechanisms) significantly affect the design choices in the implementation of a recommender system.

2.1 Collaborative and content-based filtering

The original recommendation approach that paved the way for recommender systems research is *collaborative filtering* (CF). CF is based on the assumption that users who have agreed in the past in their subjective evaluation on observed items (as expressed through their ratings) will eventually agree in the future (Resnick et al. 1994). Given the target user's ratings, the idea is to trace relationships or similarities between the target user and the remaining of the users in the database. The ratings on the target item provided by the "similar" users are summarized and directed to the target user. CF is characterized by its independence from item features which makes it applicable to almost any type of content.

Recommender systems approaches were merged with content-based filtering (CBF), an information retrieval technique that makes predictions upon the assumption

that a user's previous preferences or interests are reliable indicators for his/her future behavior. CBF performs a selection of items relevant to the ones that the user has found interesting in the past and therefore requires the analysis of the content into features. CBF is typically applied upon text-based documents or in domains with structured data (Balabanovic and Shoham 1997; Pazzani 1999). For example, CBF has been utilized in book recommendation tasks (Mooney and Roy 2000), using features such as title, author, or theme, and in Web-page recommendations (Pazzani 1999) where the more informative words are extracted (using the tf-idf weight) and utilized as features.

Collaborative filtering presents a number of advantages over CBF which make it a suitable filtering approach for several domains. CF enables the filtering of any type of content (such as videos, music, or advertisements) that cannot be analyzed in features by automated processes (Balabanovic and Shoham 1997). Even in content types where the analysis into features is feasible, content-based predictions cannot reflect the quality and taste (Herlocker et al. 1999), authoritativeness or respectfulness (Resnick et al. 1994), or "aesthetic quality" (Balabanovic 1997) of the item whenever this is necessary. Furthermore, CBF restricts the spectrum of recommendations within the boundaries of the user's current interests (Balabanovic and Shoham 1997). CF can provide recommendations concerning content that the user may have not considered in the past but has been found interesting by "similar" users. In terms of predictive performance, CF has been shown to produce more accurate predictions in the movie recommendation domain (Alspector et al. 1997; Basu et al. 1998). However, it must be noted that in domains with well-structured content, CBF undoubtedly provides useful recommendations in particular when the user presents an idiosyncratic behavior and therefore "similar" users may not exist (Smyth and Cotter 2000).

Content-based and CF approaches rely upon some form of user's expression of interest on items or item features. However, such interaction data may not always be available as for example at the initiation of the system usage where no interaction has occurred. Therefore, for such cases it may be necessary to separate the prediction process from the availability of user-driven interaction data and utilize existing knowledge in the domain exploiting other sources of data. A *knowledge-based* (KB) approach, may exploit knowledge concerning the item features, for example that product "x" belongs in a certain product category, functional knowledge concerning the mapping between a user's need and item(s) that may satisfy this need, or user knowledge. An example of a KB approach is the restaurant recommender EntreeC proposed by Burke (2002), which also combines CF for recommendations. In this system, a user may submit restaurants that he is familiar with or a set of criteria and the system returns similar restaurants. A semantic network contains a number of "cuisines", which is one of the content features. Relevant restaurants are returned according to their inverse distance from the user-defined restaurant. KB recommender systems remain rare in the field of recommender systems, mainly because of the need of knowledge acquisition, which introduces an additional complexity in the design of a recommender system. KB approaches, which have been extensively used in user modeling applications, can be marginally classified as a recommendation technique due to the additional complexity and exploitation of additional data sources (other than ratings) required.

2.2 Hybrid approaches

In order to exploit the advantages of available recommendation methods several hybrid approaches have been proposed, in their vast majority concerning combinations of CBF and CF (Balabanovic 1997; Claypool et al. 1999; Cotter and Smyth 2000; Schwab et al. 2000), or extending the two methods by demographics-based predictions (Pazzani 1999), while few of them utilize KB techniques (e.g., Burke 2002). A significant part of research in hybrid recommender systems concerns the techniques that can be used to combine the approaches since they may significantly affect the prediction outcome.

Burke (2002) classifies hybridization techniques into seven classes: *weighted* where each of the recommendation approaches makes predictions which are then combined into a single prediction; *switching* where one of the recommendation techniques is selected to make the prediction when certain criteria are met; *mixed* in which predictions from each of the recommendation techniques are presented to the user; *feature combination* where a single prediction algorithm is provided with features from different recommendation techniques; *cascade* where the output from one recommendation technique is refined by another; *feature augmentation* where the output from one recommendation technique is fed to another, and *meta-level* in which the entire model produced by one recommendation technique is utilized by another.

Each of the individual recommendation methods discussed above (CBF, CF, KB) can prove useful under certain conditions while their combinations may exploit their individual advantages. Among these methods, CF provides a reliable method to serve as a platform for the development of a personalization approach, which can be extended by other approaches optimizing the overall performance. In the following section, we review and compare CF methods and techniques, aiming to develop new methods that address their limitations and manage the effect on their predictive performance.

2.3 Collaborative filtering approaches

In CF, users are profiled by their ratings on the available items and can be represented by a *user* $\times$ *item* table, where each cell contains a user's rating on an item or is blank if the user has not observed or has not provided his/her rating for the specific item. User ratings can be collected either implicitly or explicitly. Implicit acquisition methods include the monitoring of user's interactive behavior, such as the browsing activities, page viewing time, and so on. Explicit acquisition methods refer to the direct request for provision of ratings. For example, Amazon.com^{TM} (www.amazon.com) and citeseer.org (www.citeseer.comp.nus.edu.sg/cs) request users to provide their ratings on read books or papers/articles on a one-to-five numerical scale. Ratings can also be requested explicitly upon icons [e.g., "thumbs-up"/"thumbs-down" or "smiling faces" as in Syskill and Webert (Pazzani and Billsus 1997)]. Besides the above ratings-based profiling, more sophisticated techniques may handle explicit qualitative user preferences for the development of user models (Domshlak and Joachims 2006).

Collaborative filtering approaches can be distinguished into two major classes: memory-based and model-based (Breese et al. 1998). Memory-based approaches operate upon the entire database of users in order to find the most similar to the target user and weight their recommendation according to their similarities.

The fundamental algorithm of the memory-based class is the nearest–neighbor (NN), which can be divided into three steps (Resnick et al. 1994):

(a) Measurement of similarities between the target and the remaining users in the database. Several similarity measures can be utilized to trace relationships between users, such as Spearman rank correlation (where similarities are computed upon rankings rather than ratings), mean squared difference (dissimilarity measure), or cosine vector similarity. However, Pearson correlation coefficient is typically used since it was empirically found to produce more accurate, or in the worst case equivalent results (Breese et al. 1998; Herlocker et al. 1999, 2002).
(b) Selection of the neighbors (most similar users) who will serve as recommenders. Two techniques have been employed for neighborhood selection: the *threshold-based* selection (Shardanand and Maes 1995), where users whose similarity exceeds a certain threshold value are considered as the neighbors of the target user, and the *top-n* technique in which a predefined number of *n* best neighbors is selected (Resnick et al. 1994). According to Herlocker et al. (2004), the threshold-based selection with threshold value equal to zero has been shown to preserve high levels of accuracy and coverage (the number of items for which prediction can be made).
(c) Prediction based on the weighted average of the neighbors' ratings, weighted by their similarity to the target user.

The main advantage of the NN algorithm is that it can incorporate the most recent data since the above process is performed upon a request for prediction. Therefore, the NN algorithm is mostly suitable for domains where user preferences or interests change rapidly, or with continuous updates in the available items. However, the computational cost increases at prediction time as all users are examined for their relationship to the target user. In order to deal with the scalability problem some form of heuristics that select only a subset of the users (Hill et al. 1995; Schafer et al. 2001) or other dimensionality reduction techniques can be employed.

In contrast to memory-based algorithms, model-based approaches build a model that generalizes the relationships between users or items and when prediction is requested apply the model to the target user's data. Representative model-based approaches include *clustering, dimensionality reduction,* and *classification* methods. Clustering aims at grouping users into clusters in order to exploit common behavior within clusters (Aggarwal et al. 1999; Breese et al. 1998; Pennock et al. 2000). Dimensionality reduction methods cluster users (Goldberg et al. 2001) or both users and items (Ungar and Foster 1998; Hoffman and Puzicha 1999; Hoffmann 2004) in order to reduce dimensionality and improve performance. Classification methods aim at classifying users into either of two classes labeled, for example, "like" and "dislike" (Basu et al. 1998; Billsus and Pazzani 1998; Breese et al. 1998). Other approaches trace *item-to-item* relationships–instead of user-to-user relationships–and create a model for the recommendation of items similar to the target one (Sarwar et al. 2000, 2001).

2.3.1 Comparison of collaborative filtering algorithms

Nearest–neighbor CF is "*generally accepted to be the most effective mechanism*" (Good et al. 1999) and can serve as a suitable and reliable base algorithm for the recommendation task. As Hoffmann (2004) explains "*memory-based methods have reached this*

level of popularity, because they are simple and intuitive on a conceptual level while avoiding the complications of a potentially expensive model-building stage."

The advantages of NN algorithms can be analyzed on several dimensions:

(a) They are fairly accurate. Following the review presented above, they are more accurate than most of the model-based approaches (Herlocker et al. 2002). This has been confirmed by empirical findings comparing NN algorithms to Bayesian modeling methods, in particular for non-binary ratings (Breese et al. 1998), association rules (Sarwar et al. 2001) or other clustering methods (Schafer et al. 2001) as well as classification methods (Basu et al. 1998; Good et al. 1999).
(b) They are intuitive at a conceptual level (Hoffmann 2004), easily analyzed and are the standard benchmark for the evaluation of other approaches (O'Mahony et al. 2002).
(c) In computational terms, they are a robust choice for the recommendation task (Middleton 2002), they are able to accommodate noisy training examples (Mitchell 1997), new data can be added easily and incrementally (Pennock et al. 2000), while they are capable of incorporating the most up-to date information concerning the user preferences (Schafer et al. 2001), in contrast to model-based approaches which need to rebuild the entire model when new data are introduced into the system (new users or new interaction data). In addition, they can be fairly accurate with a few training examples (Webb et al. 2001; Burke 2002).

The above advantages render the NN algorithm as an attractive CF approach. However, it inherits the intrinsic limitations of CF algorithms based on their founding principle to exploit like-minded users' opinions in the recommendation task, as discussed in the next section.

2.3.2 Limitations of CF algorithms

The most important drawback in CF algorithms is the *sparsity problem*, which refers to the low ratio of rated items to the total of available items. In general, recommender systems users rate only a small fraction of the available items, since they are not willing to invest time and effort to rate items (Aggarwal et al. 1999). Even in systems where ratings are collected implicitly, the vast amount of available items and the requirement that users have actually observed and reviewed an item makes the collection of a sufficient number of ratings a hard task. For example, sparsity levels at the MSWeb site dataset which accounts visits in various Microsoft pages is 98.4%, at the Nielsen's TV network viewing dataset 95.1% and at the EachMovie movie recommendation system 97.1% (Breese et al. 1998). The purchase data of the e-commerce site Fingerhut Inc., present 99.9% sparsity level (Sarwar et al. 2000) and the research movie recommendation system MovieLens 93.6% (Resnick et al. 1994; Herlocker et al. 2002).

In model-based algorithms, the sparsity problem affects the reliability and accuracy of prediction since the model is built upon few data points, while some model-based algorithms cannot operate on missing data. In NN algorithms, significantly affects the measurement of similarities, which is the most important step in the prediction process (Sarwar et al. 2001). It is also important to underline that sparsity refers to the entire database and can also affect the prediction on users who have rated a sufficient number of items. Indeed, if the sparsity level in the entire database is high, then few

items would have been rated in common with the remaining of the users. This problem is also significant at the initial stages of use, where very few items have been rated by the users.

Two other problems are related to the number of ratings provided by the users, both underlying the inability of CF systems to operate in *"cold start"* situations. The *new item* or *first-rater* problem, that occurs when a new item is introduced in the database, which has not been rated before by any of the users. Then CF algorithms fail to make a prediction. The second problem is the *new user* problem that occurs when a new user is introduced in the system. Since no ratings for the specific user are available, similarities cannot be computed and prediction cannot be made. In particular at the initiation of the system both new user and new item problems occur and any algorithm based on user ratings fails to make predictions.

On the other hand, CBF does not directly suffer from the sparsity effect, since predictions are made independently from the number of ratings provided by the remaining of the users in the database (besides the target user). However, the quality of prediction is affected by the number of items rated by the target user while CBF cannot make prediction for new users introduced to the system (new user problem).

Combined predictions based on both CF and CBF may partially address the sparsity problem, though they still fail to operate when the "new user" problem occurs and certainly at the initiation phase of the system. In such cases, a combination of CF/CBF and a rating-independent approach (such as KB) may prove efficient. Also, clustering and dimensionality reduction approaches aim to manage the sparsity effect by grouping users and/or items in dense subspace of the user $\times$ item matrix.

It must be noted that the development and the applicability of an integrated approach that deals with the above problems depends on the application domain and the ability to analyze content into features (in order to apply CBF), the existence of functional knowledge (in order to apply KB), and the ability to collect sufficient interaction data (in order to apply CF).

In the following section, we introduce the concept of lifestyle as a user characteristic that can influence the development of a recommendation approach that manages the sparsity problem.

3 Exploiting lifestyle segmentation methods

To address the limitations described above, we turn our focus on traditional marketing and consumer behavior theory to identify concepts and personalization approaches that remain unexploited in the recommender systems literature.

3.1 Lifestyle segmentation

In marketing theory and practice products or services are targeted to consumers by applying *target marketing* techniques (Belch and Belch 1995) represented by the *segmentation-targeting-positioning* (STP) process (Kotler 1994). Following the STP process, marketers first divide the market into homogeneous groups of consumers (*market segmentation*), select one or more appropriate market segments that best serve their objective (*targeting*), and decide the strategy to position the product in the selected segments (*positioning*) (Kara and Kaynak 1997). Markets can be segmented

on different bases such as geographic, demographic, socioeconomic, or behavioral attributes (Gunter and Furnham 1992).

One of the most effective and popular segmentation methods is *lifestyle segmentation* (Vyncke 2002), which groups consumers according to their *lifestyles*. Lifestyle is defined as the patterns in which people live and spend their time and money (Gunter and Furnham 1992). It represents the central notion in the Consumer Behavior Model (Hawkins et al. 1998) which suggests that consumers' actual and desired lifestyle (i.e., the way they would like to think and feel about themselves) are translated into daily behaviors including purchase and consumption behavior. Lifestyle is affected by a number of external (culture, subculture, demographics, social status, reference groups, family, and marketing activities) and internal factors (perception, learning, memory, motives, personality, emotions, and attitudes). Lifestyle can be quantified through psychographic research that measures constructs revealing attitudes, values and beliefs, interests and activities, demographics, media consumption, and product usage rates. The measurement of these constructs and the application of clustering techniques upon these data lead to the lifestyle segmentation. The clustering process also provides a set of classification rules, which can be applied to consumers' demographic and media consumption data to classify them into the lifestyle segments. Subsequently, the product usage rates attached to the description of the segments are used to infer the preferences of consumers and target products accordingly. A number of consumer research companies and advertising agencies have performed general lifestyle studies but the "most popular psychographic research" (Hawkins et al. 1998) is VALS, which has been developed in 1978 by SRI International (http://www.sri-bi.com/vals) and revised to VALS2 in 1989. It divides the American population into eight distinct value and lifestyle patterns, in other words, groups (or segments) of people with homogeneous lifestyle behavior.

Besides marketing theory and practice, lifestyle data have been also exploited in related fields such as Customer Relationships Management (CRM). The combination of transactional data and additional data (such as lifestyle data) from external sources has been used in CRM for the extraction of business knowledge through analytical techniques (analytical CRM) such as OLAP and Data Mining (Arndt and Gersten 2001). However, lifestyle data exploitation in CRM mainly concerns the analysis of user characteristics, behavior, and needs providing support to operational activities rather than performing automated recommendations in on-line environments.

3.2 Lifestyle in personalization research

In personalization research, personality traits (such as lifestyle) have been acknowledged as potential personalization factors (Brusilovsky 2001), but lifestyle has not been adequately studied to date, except in a few cases, such as SeAN (Ardissono et al. 2001) and Lifestyle Finder (Krulwich 1997). However, both systems limit the exploitation of lifestyle to the classification of users in lifestyle segments. In a commercial setting, one of the well-known cases is the Angara company (no longer in existence) that delivered anonymous personalization by classifying Web site visitors into lifestyle segments without requiring the monitoring of user's interactive behavior on the specific site. The company has been delivering personalized content even to first-time visitors by retrieving user-related data from cookies previously placed by one of the company's data partners. This solution is extremely practical and applicable for first-time visitors of a Web site but it also requires classificatory data (monitored

by the cookies) and the maintenance of a huge database of profiles (in fact Angara maintained a database of 150 million anonymous profiles).

In research indirectly involving personality or behavioral factors, Pennock et al. (2000) propose a Personality Diagnosis (PD) algorithm, based on the assumption that there is an association between how people rate items and their personality type (modeled as a latent variable). Other approaches aim at clustering users assuming that behavioral relationships exist among them (Breese et al. 1998; Ungar and Foster 1998). Besides the fact that such approaches are not concerned with the notion of lifestyle per se, another major difference is that they rely upon available data and are therefore affected by the sparsity problem, while the personality factor upon which clusters are developed is not specified.

In contrast, lifestyle is a meaningful behavioral predictor that can group users of recommender systems concerning the filtering of a wide range of products and services. In order to exploit the lifestyle factor we propose a number of personalization algorithms, which are presented in the remainder of the paper in an incremental level of personalization (segment-level, user-level, item-level), as follows:

First, a *segment-level* personalization approach called "segmentation-based" is proposed that refers to the classification of the target user into a predefined segment exploiting existing knowledge. A typical example of related work is the stereotypical approach (Rich 1979, 1983) where users are classified into stereotypes and inferences about his/her future behavior are drawn from the description of the stereotype itself.

Second, a set of *user-level* approaches are proposed (the "lifestyle," "hybrid," and "integrated") whose predictive performance improves compared to each other. Algorithms in this class are based on the dynamic development of a "personal" neighborhood for each user. For example NN or clustering approaches (Breese et al. 1998) belong in this class.

Third, an *item-level* approach is proposed, called "best-item," where the target item is also taken into account. For example, item-to-item approaches or approaches that cluster both users and items (Ungar and Foster 1998) take into account the similar items to the target one. Besides CF, content-based approaches consider the features that describe an item in order to make a recommendation and therefore can be classified as item-level.

4 Segment-level personalization

The development of target user's neighborhood based on similarities computed upon few ratings may lead to the erroneous selection of actually "bad" neighbors as "good" ones and vice versa. The idea underlying the use of lifestyle is that the above risk may be avoided by developing for each target user a "lifestyle" neighborhood. Since people can be discriminated upon their lifestyles (Chaney 1996), and consumers found in the same lifestyle segment present similar behavior, the members of a "lifestyle" neighborhood can be considered as reliable recommenders to each other. As a result, the identification of similar users in terms of their lifestyle, will restrict the search space to users that present this form of similarity (lifestyle), avoiding the effects of misleading similarity computations. A direct implementation of this reasoning is to classify users into existing lifestyle segments.

More specifically, a *segmentation-based* approach is proposed that can be divided into a classification and prediction step, described below.

4.1 Classification step

In traditional marketing approaches users can be classified into lifestyle segments on the basis of psychographic variables measured through appropriate questionnaires. However, forcing all users of a personalized system to fill in such questionnaires is rather difficult and annoying (Balabanovic and Shoham 1997; Kobsa et al. 2001). On the other hand, lifestyle segments are typically described by some form of static data that change infrequently, such as demographics and/or media consumption data (Hawkins et al. 1998), which also serve the role of classificatory data. Thus, a straightforward approach is to utilize the above users' data for the classification task depending on their availability in the application domain.

An alternative approach is based on the assumption that members of a lifestyle segment present similar behavior within the segment and differentiated behavior between the segments and therefore their behavioral data (ratings) may be utilized for the classification process. In order to develop the classification model for this classification approach we need a labeled training set consisting of users with known membership in the lifestyle segments. This can be achieved by having a portion of the population fill-in the psychographic questionnaire and classified accordingly. The sample's on-line behavior (ratings) is coupled with their respective labels and a learning algorithm is employed to produce the classification rules. The rules are then continuously applied to the above interaction data for each individual, so that the segment to which the user belongs can be dynamically determined and re-assessed if needed. As the amount of data that are being monitored for each user increases (through usage), updated classification rules are developed and applied, thus adjusting the classification into clusters.

Each of the above classification approaches presents its own advantages/disadvantages. The first one (based on static data) is completely independent from the availability of behavioral data (ratings) but requires the collection of users' static data. The second classification approach (based on ratings) presents the advantage of exploiting lifestyle segments without requiring the collection of additional data (besides ratings) from each user but only psychographic data from a portion of the population. The selection of the appropriate approach depends on the amount and type of data available in the application domain.

4.2 Prediction step

Assuming the classification of users into lifestyle segments, the prediction task can rely on the assumption that lifestyle segments represent well-discriminated clusters of users (a basic assumption for the development of the clusters). Then, two prediction strategies can be employed:

- *Expert-driven prediction*, which directly exploits the classification of users into lifestyle segments. Indeed, the human expert (marketer) pre-assigns stereotypical behavior to the members of each segment (based on marketing data and/or experience), exploiting the acquisition of the class label (lifestyle segment membership) of the users.
- *Center-based prediction*. A more dynamic type of prediction utilizes the behavior of segment members to infer a characteristic and representative behavioral pattern for each segment. The central tendency (also known as the centroid) for each

segment can be easily formulated by aggregating all users' preferences in a single vector by averaging the ratings assigned to each item by all users. In the binary ratings case, this can be implemented by selecting the most frequently observed class ('0' or '1') for each item, while in the numerical ratings case the center of the segment is the vector containing the mean rating for each item available (Hair et al. 1998)

Both prediction strategies work at the cluster (segment) level and depend on the quality of the lifestyle segments as well as on the expert's quality (expert-based prediction). Their main difference is that the expert-based prediction does not rely on the ratings of the segment members, while the opposite holds true for the center-based approach. In addition, the utilization of the first classification approach described in the previous section (based on static data) and the application of an expert-based prediction is completely independent from rating availability and therefore it addresses the cold start problem. Empirical results from previous research indicate that the expert-based approach described above outperforms the base case of non-personalized recommendations while the center-based prediction outperforms the expert-based prediction for both numerical and binary ratings with Bayesian Networks used for the classification task (Lekakos and Giaglis 2004).

The segmentation-based prediction described above may be used at the initial phases of a recommendation system and refined through the application of a more dynamic prediction strategy.

4.3 Limitations of the segment-level personalization

The utilization of existing lifestyle segmentation methods leads to a number of limitations that affect the performance of the segmentation-based approach:

(a) The isolation of the prediction process within the lifestyle segments may ignore behaviorally similar users that can be found in other segments.
(b) The proprietary nature of lifestyle segments. Lifestyle segments are developed by marketing/consumer research companies or organizations that withhold their rights of use. As a consequence, this restricts the accessibility to detailed data for further elaboration or validation (Gunter and Furnham 1992; Beatty et al. 1998; Mowen and Minor 1998). Although lifestyle segments have proven very useful in marketing practice, they cannot be fully explored due to the limited availability of relevant data and information. For example, in the segmentation-based approach, we would have gained much in prediction accuracy if we were able to extract from raw data smaller lifestyle groups (clusters) rather than the "large" lifestyle segments that are commercially available.
(c) The sparsity problem affects the performance of the segmentation-based approach, when ratings are used for the classification of users in the segments, affecting the classification process both at the training phase (development of classification rules) and at the application of the rules on sparse datasets.

Separating the personalization process from the use of lifestyle segments and taking into account the closeness or similarity between the target and the remaining of the users (ignored in the segment-level approach) would lead to more dynamic and eventually more accurate predictions.

5 User-level personalization

In order to achieve user-level personalization avoiding the classification of users into static lifestyle segments, we measure "lifestyle" similarities directly among all available users (in a CF manner) and develop a dynamic "personal" neighborhood for each user. Similarity measurement should be performed on the basis of some type of suitable user data that expresses the user's lifestyle. The identification of the "lifestyle" data that may serve our objective is discussed in the next section followed by the presentation of three "user-level" personalization approaches. The first one (called "lifestyle") avoids the direct use of lifestyle segments, and serves as the basis for two improved versions (called "hybrid" and "integrated," respectively) that manage the sparsity effect by combining the advantages of the lifestyle and the NN approaches.

5.1 Lifestyle data

The lifestyle data that will be used for the measurement of similarities should be lifestyle indicators, i.e., they should be associated with the membership of a user into a lifestyle segment. This expresses our confidence in the marketing concept of lifestyle as a predictor of human behavior. They should also be independent from the availability of ratings in order to reduce the sparsity effect and sould be easily collectable in the application domain.

Previous empirical findings suggest that user demographics (Krulwich 1997) in combination with other user data such as hobbies (Ardissono et al. 2001) or customer credit data in the banking domain (Peltier et al. 2002) may be used for the classification into lifestyle groups. Other empirical findings in the domain of interactive television advertisements suggest that demographics and television program preferences data are significant indicators of a user's lifestyle (Lekakos and Giaglis 2005) while they satisfy the second requirement above. Furthermore, they can be easily collected in the domain of interactive television either off-line (at the subscription to the service) or—in the case of demographics—through the use of electronic forms provided by the personalized system (Bozios et al. 2001). Television program preferences may be monitored through the set-top box (provided that we also know who is watching).

In order to confirm the significance of the above attributes as lifestyle indicators we performed statistical analysis (multinomial logistic regression) on a sample of 502 consumers (with known lifestyle segments). The extracted lifestyle attributes are the demographics "age," "marital status," and "education" along with the consumer's preferences on eight program genres: "Documentaries," "Cartoons," "Football/basketball/volleyball games," "Video clips," "Domestic comedy series," "Discussions/interviews," and "News." These attributes have been measured (among others) in a psychographic questionnaire (answered by the sample of 502 consumers) used by the local subsidiary of the multi-national consumer research company AGB for the classification of consumers into the lifestyle segments. The above attributes meet our requirements as lifestyle indicators and they will be used in the development of lifestyle approach where their effectiveness will be ultimately evaluated.

5.2 The lifestyle approach

The above attributes are encoded as binary variables valued either "0" when a user has stated that he/she does not like the specific program category or as "1" if a user has

stated that he/she likes it. Encoding uniformly the above attributes as binary variables we are able to profile each user by attribute-value pairs and compute "lifestyle" similarities directly among the users. While several similarity measures can be applied on binary variables the Pearson correlation coefficient is selected since it has been used in the measurement of similarities upon demographic data (Pazzani 1999) as well as for consistency reasons with the NN algorithm that will be used for the evaluation of the proposed approach. The "*lifestyle*" approach (Lekakos and Giaglis 2006), is described by the following steps:

1. Measure similarities between the target and the remaining users based upon data associated with their lifestyle by applying the Pearson correlation coefficient formula (1):

$$w(i,j) = \frac{\sum_k (I_{i,k} - \overline{I_i})(I_{j,k} - \overline{I_j})}{\sqrt{\sum_k \left(I_{i,k} - \overline{I_i}\right)^2 \sum_k \left(I_{j,k} - \overline{I_j}\right)^2}}, \quad (1)$$

   where $I_{i,k}$ and $I_{j,k}$, refer to k^{th} lifestyle indicator available in common for the i^{th} (target user) and j^{th} users, and $\overline{Ii}$ and $\overline{Ij}$ to the corresponding means.
2. Formulate the target user's neighborhood, based on the similarity measures described in step 1, by selecting users who score above a certain threshold.
3. Predict the target user's rating on the target item by aggregating lifestyle neighbors' ratings weighted by the lifestyle similarities developed at step 1. Aggregate the target user's preferences into a prediction for the target item by applying Eq. 2 (Resnick et al. 1994; Hill et al. 1995; Shardanand and Maes 1995):

$$R_{i,p} = \overline{R_i} + \frac{\sum_{j=1}^{m} w(i,j)(R_{j,p} - \overline{R_j})}{\sum_{j=1}^{m} |w(i,j)|}, \quad (2)$$

where–identically to the prediction formula of the NN approach–$R_{i,p}$ is the rating to be predicted for user i and for item p, $\overline{R_i}$ is the mean of the ratings of user i for all items that user has provided his/her ratings, the weight $w(i,j)$ is the similarity measure between user i and j and $R_{j,p}$ is the rating of user j for item p and $\overline{R_j}$ is the mean of ratings of user j in a neighborhood of size m.

The lifestyle approach is independent from the availability of ratings at the most important step of the above process (Sarwar et al. 2001) where similarities are computed. However, in order to make predictions at the third step above a number of ratings are required, similarly to the NN algorithm. It can be easily observed that the main difference between the lifestyle and the Pearson-based NN approach (that we will refer to as *Pearson-based* hereafter) is at the first step of the proposed method where lifestyle indicators, instead of ratings, are used for the computation of similarities. This step results into different neighbors and different weights indicating the importance of each neighbor's rating utilized in the prediction formula. Thus, the prediction accuracy of the proposed approach depends of the validity of the hypothesis that lifestyle neighbors are more reliable than ratings-based neighbors in sparsity conditions. The validity of this hypothesis will be examined in the next sub-section.

5.2.1 Empirical evaluation of the lifestyle approach

The objective of the empirical evaluation is to measure the predictive accuracy of the lifestyle approach in comparison to the Pearson-based approach at different sparsity levels. In order to avoid underestimating the predictive performance of the Pearson-based approach for its comparison to the lifestyle approach, we carefully tuned the algorithm's parameters. Besides the first and third steps of the Pearson-based prediction process where standard formulas (Pearson correlation and prediction formula (2)) are applied, the selection of the threshold value in the second step of the prediction process is a crucial parameter for the development of the target user's neighborhood. For this choice we followed the suggestions by Herlocker et al. (2004) using (in both approaches) a threshold value equal to zero and we confirmed this suggestion through experimentation on the dataset used in our empirical research. The two algorithms (lifestyle and Pearson-based) differ only in the use of the "lifestyle" data following the same algorithmic reasoning, implementation, and evaluation and therefore the differences in their performance can be attributed in the use of these data.

The sample used in our experiment consists of 37 individuals drawn from our research group, including academic (19%), research (73%), and technical staff (8%), consisting of 62.2% males and 37.8% females, aged 18–24 (10.8%), 25–34 (67.6%), 35–44 (18.9%), and 45–54 (2.7%). The users were shown 65 advertisements selected from seven product categories (food and drink, fast moving consumer goods, computer and technology, family and home, books and magazines, public services, finance and investment, and autos). Users were asked to provide their overall evaluation for each advertisement in a form of a rating in a one-to-five scale. It must be noted that an advertisement may be liked because of the creative part, the featured actor, the music theme, its entertaining nature, or because the consumer is interested in the advertised product (similarly to the movie domain where users might favor a movie due to the theme, the actors, the director, etc.). However, in CF we are interested in a rating as an expression of the user's "overall taste" on an item independently from the factors that may affect this subjective evaluation (Harter 1996 showed that as many as 80 factors may affect this evaluation). Finally, participants filled-in a questionnaire providing their demographic and TV program preferences data as required by the lifestyle algorithm.

Both algorithms are tested upon the same set of users under a leave-one-out cross-validation technique, which is the recommended technique for the comparative evaluation of learning algorithms upon small samples (Cawley and Talbot 2003). This method replicates the error estimation process n times for a sample of size n by considering each user in the original sample as the test set (target user) and the remaining sample of size $n - 1$ as the training set.

In order to manage the sparsity effect, a certain number of randomly selected ratings are removed following the experimental design for the empirical analysis of CF algorithms introduced by Breese et al. (1998) who describe a set of experimental protocols, called *Given* 2, *Given* 5, *Given* 10, and "All-but-one." The *Given n* protocol involves the random selection of 2, 5, or 10 votes (corresponding to "*n*") from each test (target) user as the observed ratings, which are then used to predict the remaining ratings. The observed ratings are indicated as the *training set of items* and the ratings to be predicted as the *test set of items* for each target user. The various "given" protocols examine the performance of the algorithms when relatively little is known about the target user. The "All-but-one" protocol measures the performance of the

Table 1 Overall performance of the algorithms

	Given 2	Given 5	Given 10
Lifestyle	1.1639	0.8265	0.7850
Pearson	1.1857	0.8416	0.7942
t-value	−5.917	−3.818	−1.946
(*p*)	(<0.0001)	(0.000)	(0.054)

algorithms when as many as all-but-one ratings are considered available, representing in fact a non-sparsity condition and therefore it is beyond the scope of our empirical evaluation.

The algorithms' prediction error is measured using the *Mean Absolute Error* (MAE), which is the average difference between the predicted and the actual rating value and is commonly used for evaluating the predictive performance on numerical ratings (Shardanand and Maes 1995; Breese et al. 1998; Claypool et al. 1999; Herlocker et al. 2002; Melville et al. 2002). We are interested in predicting a rating for all advertisements rather than selecting the top-*n* of them because this may lead to the exclusion of certain advertisements (that may be prioritized by other factors such as campaign objectives or specific market conditions). Therefore, MAE is considered as an appropriate measure for our evaluation task (Herlocker et al. 2004). Differences in MAEs are compared using paired *t*-tests (in all cases presented below the normality requirement is met).

In order to cross-validate the results we replicated the experiment for each protocol and the averaged results are depicted in Table 1.

The above results confirm that *lifestyle* approach gives lower error levels at the Given 2 and 5 protocols at 95% significance level and at 90% significance level for the Given 10 protocol. It must be noted that the 90% confidence level has been statistically used for comparison purposes by Breese et al. (1998).

The *lifestyle* approach—although significantly better—is also affected by the number of available items due to the mean rating value that is used in the prediction formula (in both the *lifestyle* and the Pearson-based approach). It is straightforward that the mean value of very few ratings cannot reliably depict the actual mean value as it would have been derived from several ratings. The mean value of the available ratings for the target user significantly affects the final prediction, since it serves as the estimation of the neutral rating of the target user.

5.3 A hybrid approach

In this section, we propose a *hybrid* recommendation mechanism that utilizes the *lifestyle* approach to make predictions upon the (eventually few) available ratings for the items unobserved by the target user, populating his/her rating vector. A similar–in principle–approach has been followed by Good et al. (1999) in the movie domain, where personal information filtering (IF) agents acting as virtual users make predictions on the basis of item features (rather than user-related lifestyle data). An analogous approach has been proposed by Sarwar et al. (1998) in the news domain where a single rating agent evaluates the quality of news articles participating in a CF recommendation.

In the proposed hybrid approach, a new user representation is introduced, called "*pseudo-user*" (Melville et al. 2002), consisting of the original ratings provided by

the target user and the ratings predicted by the application of the *lifestyle* approach. The substitution of the original target user by the pseudo-user leads to an increase in the number of overlapping ratings with the remaining of the users in the database. In addition, the pseudo-user contributes in the final prediction for any target item (i.e., any of the target user's missing ratings) by the rating predicted by the lifestyle approach. The target user and his/her corresponding pseudo-user are perfectly correlated (correlation coefficient = 1) and therefore we use an amplified weight (>1) to strengthen the pseudo-user's contribution in the prediction (underlying the fact that the pseudo-user is the most "similar" to the target user in the database).

More specifically, if the target user t has provided his/her ratings for k items $\{R_{t,1}, R_{t,2}, \ldots, R_{t,k}\}$ in a total of n items, then the steps of the approach can be described as follows:

Lifestyle Prediction

1. Measure similarities between the target and the remaining users upon lifestyle data.
2. Formulate the target user's neighborhood, by incorporating all users whose similarity's level exceeds a certain threshold (typically threshold = 0).
3. Predict the $n - k$ ratings $\{L_{t,1}, L_{t,2}, \ldots, L_{t,n-k}\}$ for the target user using Eq. 2 weighting the contribution of each user according to his/her similarity to the target user.

Pseudo-user formulation

4. Introduce in the target user's neighborhood a new user, the pseudo-user, whose rating vector is defined as follows:
$$R_{t,i} = \begin{cases} R_{t,i}, & \text{if rating for item } i \text{ has been provided by the user,} \\ L_{t,i}, & \text{if item } i \text{ has not been rated by the user.} \end{cases}$$

Pearson-based prediction

5. Measure similarities between the pseudo-user and the remaining of the users using the Pearson correlation. Assign the respective weights to each of the users and an increased weight to the pseudo-user (in order to strengthen its contribution to the final prediction).
6. Formulate the target user's neighborhood by selecting users above a certain threshold.
7. Produce the prediction by the Pearson-based prediction formula (Eq. 2) for the neighbors selected in step (6) and weights computed in step (5).

The increase in the number of overlapping ratings between the target and the remaining of the users in the database (through the introduction of the pseudo-user) leads to a more accurate estimation of similarities. In addition, users with no ratings in common with the target user that were excluded from the prediction process are now evaluated as potential neighbors. However, the final prediction depends on the amount of ratings available for the target item that remains unchanged after the application of the hybrid approach. It is clear that the notion of pseudo-user can be easily extended to all of the users in the database providing a dense *user* × *item* table which addresses both problems above, as discussed in the next section.

5.4 An integrated approach

Exploiting further the notion of pseudo-user, we extend its utilization beyond the target user. More specifically, the extended *hybrid* approach (which we will call "*integrated*" approach hereafter), exploits the prediction of the "missing" ratings for each and every user in the data set, formulating a new pseudo-user matrix rather than a single pseudo-user vector

The prediction process closely follows the steps of the *hybrid* approach described above (Sect. 5.3). More specifically, the lifestyle prediction process (steps one to three above) is repeated for each user with missing ratings, leading to completely dense pseudo-*user* × *item* matrix, which is then utilized for the Pearson-based prediction of the original target item (steps five to seven above). However, the lifestyle prediction error is transferred and eventually magnified through the application of the Pearson-based prediction process in the integrated approach. Therefore, the effect of massive substitution of all users by their respective pseudo-users remains to be investigated. The main hypothesis to be tested is that the integrated approach significantly outperforms the Pearson-based approach.

5.4.1 Empirical evaluation of the integrated approach

The experimental design utilizes once again the "Given" protocols, upon the sparse *user* × *item* matrix. Furthermore, we extend the number of given ratings beyond the Given 2, 5, and 10 ratings. Specifically, the performance of all algorithms discussed so far is measured upon 2, 5, 10, 15, 20, 25, 30, 35, and 50 ratings, since our aim as well as our expectations concerning the integrated approach is that it outperforms the lifestyle one beyond high-sparsity conditions.

In Table 2, the MAEs for each of the approaches are presented as well as the p-values (in parentheses) of the respective paired t-tests indicating the significance of differences ($p < 0.05$) between the integrated and the rest of the approaches.

The results suggest that the *integrated* approach is significantly better than the Pearson-based, as well as than all approaches examined so far (besides the Given 2 protocol). All approaches increase their performance as more items are added in the training test confirming the theory-driven expectation concerning the sparsity effect upon the performance of a learning algorithm. In contrast to the *hybrid* and *lifestyle* approaches, the *integrated* approach firmly outperforms the Pearson-based in the range of 5–35 available ratings, while the improvement is decreased for 50 ratings.

Table 2 Comparison of the *integrated* approach to the Pearson-based, *lifestyle*, and *hybrid* approaches

	Pearson	Lifestyle	Hybrid	Integrated
G2	1.1071 (0.010)	1.1053 (0.142)	1.1044 (0.185)	1.1048
G5	1.0056 (0.0001)	0.9881 (0.0001)	0.9805 (0.0001)	0.9783
G10	0.9333 (0.0001)	0.9286 (0.0001)	0.9249 (0.0001)	0.9057
G15	0.906 (0.0001)	0.9059 (0.0001)	0.9077 (0.0001)	0.8776
G20	0.8859 (0.0001)	0.8884 (0.0001)	0.8894 (0.0001)	0.8556
G25	0.8679 (0.0001)	0.8653 (0.0001)	0.8713 (0.0001)	0.8309
G35	0.8352 (0.0001)	0.8381 (0.0001)	0.8437 (0.0001)	0.8081
G50	0.7784 (0.045)	0.7867 (0.006)	0.7903 (0.004)	0.7687

Table 3 Comparison of the *integrated* and Pearson-based approaches on the second dataset (*p*-values of *t*-tests in bold indicate significant differences)

Protocol	Pearson	Integrated	*p*-value
G2	1.249416	1.248323	0.69
G5	1.114178	1.08379	0.01
G7	1.086942	1.055267	0.002
G9	1.026026	1.003755	0.009
G12	0.987844	0.97229	0.011
G16	0.981885	0.963503	0.004
G23	0.920992	0.910384	0.092

In addition, a second experiment has been performed that evaluates the predictive performance of the *integrated* approach compared to the Pearson-based (Table 3). The sample in this experiment consists of 34 individuals-employees in a commercial firm. 42.2% of the sample is males and 58.8% females aged 15–34 (17.6%), 35–44 (44.1%), 45–54 (23.5%), and 55–64 (14.7%). Their studies (educational level) include high school (26.5%), higher education (35.3%), and university/postgraduate studies (38.2%). The sample was shown a set of 30 advertisements (subset of the set of the 65 advertisements) and the evaluation methodology was identical to the one used in the previous experiment. Given the fewer number of items used in this experiment, instead of considering 2, 5, 10, 15, 20, 25, 35, and 50 items available for each user we selected 2, 5, 7, 9, 12, 16, and 23 items that correspond to the sparsity levels of the previous experiment. This enables us to facilitate the interpretation of the results and compare the behavior of the integrated approach in the two experiments.

The above results indicate that the behavior of the *integrated* approach closely follows the incremental performance observed in the previous experiment outperforming the Pearson-based approach for all protocols with the exception of "given2" and "given23" protocols where no statistically significant differences are found.

The improved performance of the *integrated* approach is mainly based on the increase in the density of the ratings matrix. Thus, other implementations of the integrated approach that populate the ratings matrix may lead as well to predictions with improved accuracy. Indeed, one of the advantages of the integrated approach is that it can accommodate any algorithm (or variation of the lifestyle and/or Pearson-based algorithms) that operates upon ratings and may eventually have a positive impact on the performance of the *integrated* approach. For example, one alternative implementation of the *integrated* approach is to compute prediction using the Pearson-based approach in order to populate the ratings matrix and produce the final prediction by a second application of the Pearson-based on the dense matrix (Table 4)

Table 4 Comparison of two variations of the *integrated* approach

	Integrated (Lifestyle–Pearson)	Integrated (Pearson–Pearson)	*p*-value
G2	1.1048	1.1063	0.20930
G5	0.9783	0.9909	0.00001
G10	0.8980	0.9012	0.02185
G15	0.8776	0.8717	0.00001
G20	0.8556	0.8481	0.00001
G25	0.8310	0.8256	0.00694
G35	0.8082	0.8020	0.11239
G50	0.7691	0.7682	0.59684

The above results indicate that we can gain improvements in the accuracy—equivalent to the ones produced by the use of lifestyle data in the integrated approach. One clear advantage of using repetitively the Pearson-based algorithm is that we do not need to collect additional lifestyle data. However, using the Pearson-based algorithm significantly reduces the coverage (i.e., the number of items that prediction can be made for) of the prediction, particularly in high sparsity conditions (such in the "given 2" or in the "given 5" protocols). It can be easily observed that it is highly unlikely to find enough users with overlapping ratings (in particular in real-world systems with a vast amount of available items). In contrast, the integrated approach based on lifestyle data overcomes this problem (providing 100% coverage) since it does not have to measure similarities upon ratings.

It must be noted that the application of the *integrated* approach may increase the computational cost of the prediction process up to $O(n^2m)$, for n users and m items. However, this cost (that refers to the computation of similarities between all pairs of users), may be transferred to an off-line phase since lifestyle data change infrequently and therefore there is no practical need to perform these computations at prediction time. Thus, the cost at prediction time can be reduced to $O(nm)$ which is even less in the case of the hybrid approach ($O(m)$) since all computations refer to the target user (i.e., we develop one pseudo-user rather than all pseudo-users). In cases of very large databases dimensionality reduction methods can be incorporated in the *integrated* approach—as a preprocessing off-line step—in order to deal with large number of users and/or items. Scalability problems are typically addressed by Latent Semantic Indexing (LSI) (Sarwar et al. 2001), Principal Component Analysis (PCA) (Goldberg et al. 2001), probabilistic Latent Semantic Analysis (pLSA) (Hoffman and Puzicha 1999; Hoffmann 2004), attribute selection methods (Moore and Lee 1994), or simply by a heuristic selection of a subset of users (in the case of large number of users).

6 Item-level personalization

The *integrated* approach is designed to manage the sparsity effect exploiting the behavior the lifestyle and Pearson-based approaches: the former performs better when fewer ratings are available while the latter is more accurate as the number of available ratings increases. The integrated approach utilizes firstly the lifestyle algorithm to make the original predictions and the Pearson-based for the final prediction on the populated ratings matrix. However, as more user-driven (actual) ratings become available and consequently less lifestyle predictions are required, the sparsity effect is reduced and the predictive performance of the integrated approach approximates the performance of the Pearson-based approach. The usefulness of the lifestyle approach may be further extended beyond the boundaries of high-sparsity conditions.

6.1 The best-item approach

A careful investigation of the performance of the *lifestyle* and Pearson-based approaches throughout all protocols, on an item-by-item basis, reveals that the relative performance of the two algorithms is inconsistent with their averaged performance. For example, in low-sparsity levels where the Pearson-based outperforms the lifestyle approach on average, for certain items and users the lifestyle algorithm produces more accurate predictions.

Table 5 Performance of the "optimum" *best-item* compared to the *lifestyle*, Pearson-based, and *integrated* approaches

	Lifestyle	Pearson	Integrated	Best-item	*t*-test (Best-item/Pearson)	*t*-test (Best-item/integrated)
G2	1.1053	1.1071	1.1048	1.0980	$t = -30.369$ ($p < 0.0001$)	$t = -11.7951$ ($p < 0.0001$)
G5	0.9881	1.0056	0.9783	0.9678	$t = -24.502$ ($p < 0.0001$)	$t = -6.972$ ($p < 0.0001$)
G10	0.9286	0.9333	0.9057	0.8787	$t = -24.132$ ($p < 0.0001$)	$t = -10.764$ ($p < 0.0001$)
G15	0.9059	0.9060	0.8776	0.8658	$t = -23.347$ ($p < 0.0001$)	$t = -3.59273$ ($p = 0.004$)
G20	0.8885	0.8859	0.8556	0.8536	$t = -17.874$ ($p < 0.0001$)	$t = 0.59633$ ($p = 0.2773$)
G25	0.8653	0.8680	0.8310	0.8268	$t = -16.431$ ($p < 0.0001$)	$t = -0.8625$ ($p = 0.1970$)
G35	0.8382	0.8352	0.8082	0.7943	$t = -15.478$ ($p < 0.0001$)	$t = -2.179$ ($p = 0.018$)
G50	0.7867	0.7785	0.7688	0.7347	$t = -18.553$ ($p < 0.0001$)	$t = -6.368$ ($p < 0.0001$)

Thus assuming a hybrid approach, which we call "*best-item*" hereafter that based on the performance history of the two above algorithms (i.e., the "base" algorithms) is able to predict and apply on a specific target item the best performing one, the average accuracy is expected to improve. In order to examine the validity of this hypothesis, we measure the performance of the "optimum" best-item approach where for any given target item both lifestyle and Pearson-based predictions are computed and the more accurate one is selected (as the output of the "best-item"). The averaged prediction results (on the sample of 37 users) are compared to the performances of the approaches discussed so far (Table 5).

The results suggest a rather impressive accuracy of the *best-item* approach compared to the Pearson-based, while compared to the integrated approach it presents significantly improved performance besides the cases where 20 and 25 ratings are available. However, the most important observation concerning the behavior of the best-item approach is that it continues to increase its difference with the Pearson-based approach in low-sparsity conditions (G35 and G50 protocols), in contrast to the approaches proposed so far.

We further examined the best-item's performance on the MovieLens dataset (www.movielens.org) that contains a million ratings provided by 6,040 users for about 4,000 movies whose titles and production year are also available. Each user has rated at least 20 and on average 166 movies in the one-to-five rating scale. In the absence of lifestyle data in the MovieLens dataset, we evaluated the best-item approach using as base algorithms the Pearson-based and a content-based algorithm. The content-based algorithm is based on the computation of similarities between movies using the cosine vector similarity measure as described by Karypis (2001). Additional features for the movies (genre, cast, director, writing credits, producers, and keywords) for the content-based implementation were collected by a web crawler from the Internet Movie Database (IMDb) website (www.imdb.com). We computed content-based, Pearson-based, and best-item predictions and measured the mean absolute errors for the 20% of their ratings, using the remaining 80% as the training set (Table 6).

Table 6 *Best-item* performance on the MovieLens dataset

t-test (Pearson-based/ Best-item)	*t*-test (Content-based/ Best-item)	Content-based	Pearson-based	Best-item
111.392 ($p = 0.000$)	158.840 ($p = 0.000$)	0.914822	0.772063	0.583496

The above results support the findings of the previous experiment demonstrating that by selecting the best performing from a set of available algorithms then the overall prediction accuracy may significantly increase.

The best-item approach does not aim to produce accurate predictions by managing the sparsity problem (as done in the *integrated* approach) but by selecting the best performing algorithm for each rating to be predicted (Fig. 1). More specifically, let D be the dataset containing the ratings of n users on k items represented by an $n \times k$ table. Assume we want to predict a rating P_{um} of a target user u for a target item m (with unknown or unobserved rating). Also let $\{R_{im}\}$ be the set of ratings provided for the target item by the remaining $n-1$ users and S the set defined as $S = D - \{R_{im}\}$ (i.e., excluding all ratings for the target item) containing the sequences of ratings from all i users for the remaining $k-1$ items. We divide S into training set L and test set T. Let $L^* = L \cup \{R_{it}\}$ be the set that contains training examples $\{(R_i), R_{it}\}$ where R_i are the sequences of ratings of all users i for all items in L, and R_{it} are the ratings of all users i for a given item t in T. Then the best-item approach can be described as follows:

- *Phase* 1: Development of the "success" table
  The base algorithms (e.g., lifestyle and Pearson-based) make predictions for each of the R_{it} ratings (ignoring its actual value for the user at hand) and the predictions are compared with the actual rating. The output of this comparison is a single binary value either "0" if the Pearson-based is more accurate than the lifestyle algorithm or "1" in the opposite case. The process is repeated for all ratings in the test set T, leading to a set of meta-features represented by a *user* $\times$ *item* table—called the "success" table — which contains the sequences of successes of the lifestyle algorithm against the Pearson-based.
- *Phase* 2: Assignment of the class labels
  In the next phase we assign a class label (i.e., "0" or "1") to each sequence in the "success" table by applying the two learning algorithms for predicting the ratings R_{im} (ignoring its actual value for the user at hand) concerning the (original) target item m and measuring their relative performance. The outcome of this process is a set of training examples (sequences of 0's and 1's and their respective labels).
- *Phase* 3: Rating prediction
  At the final phase a learning algorithm (literally any supervised learning algorithm) is trained upon the set of meta-features (with known class labels as defined above). The classifier produced is applied upon the "success" sequence of target user u providing the best-performing algorithm which is applied for the prediction of P_{um}.

A key difference of the *best-item* approach with respect to the approaches presented so far is that user ratings serve not only as input to the prediction process but also as a feedback mechanism in order to evaluate the performance of the base approaches and

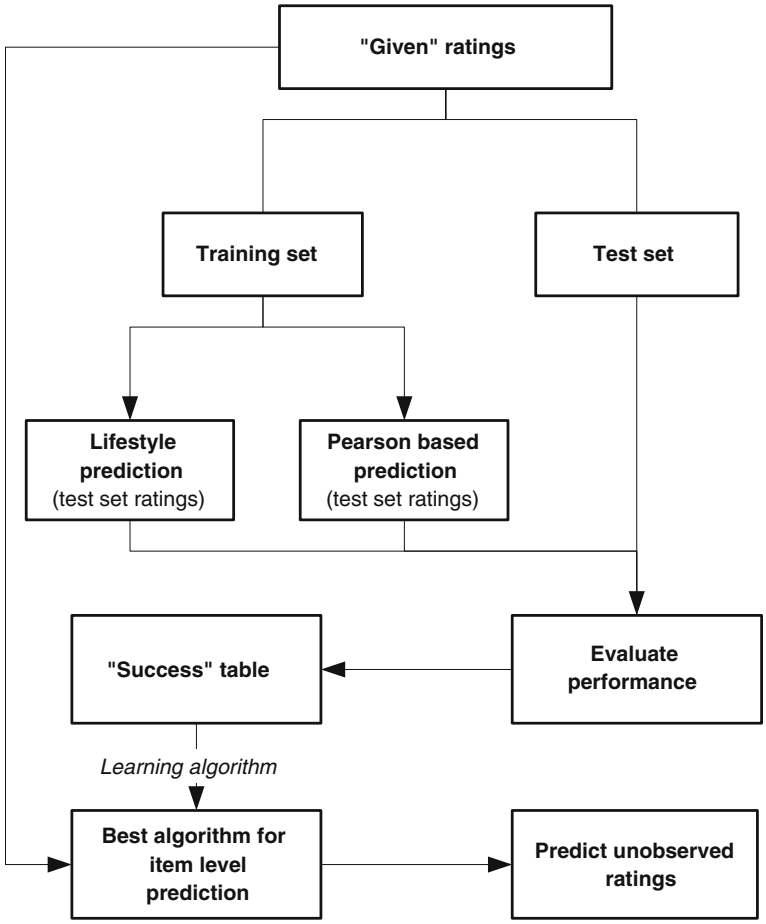

Fig. 1 *Best-item* learning and prediction process

develop the learning framework. The above formulation of the prediction problem as a learning problem enables the application of any supervised learning algorithm. In the present implementation, we employ the Naïve Bayes classification algorithm, which is computationally non-expensive (Domingos and Pazzani 1997), and has been applied in several practical problems and has been shown to perform very well compared to more complex algorithms (Mooney and Roy 2000; Melville et al. 2002). In our experimentations, we compared the Naïve Bayes and the Bayesian networks algorithms and found that Naïve Bayes performs better as described in the next section (Bayesian Networks MAE = 0.7492, Naïve Bayes MAE = 0.7380).

6.1.1 Empirical evaluation of the best-item approach

The objective of the empirical evaluation of the best-item approach is to examine whether it significantly outperforms the Pearson-based at high-density levels. Thus, we focus on the "Given 50" protocol, where we withhold 15 ratings to be predicted for

Table 7 Empirical evaluation of the Naïve Bayes best-item approach compared to the lifestyle, Pearson, and optimum best-item

Lifestyle	Pearson	Integrated	Bayesian Net "Best-item"	Optimum "Best-item"	Naïve Bayes "Best-item"
0.7647	0.7528	0.7564	0.7492	0.7115 $t = -13.128$ $(p < 0.0001)$	0.7380 $t = -4.168$ $(p < 0.0001)$

each user and utilize the remaining ratings as input for the prediction task. Moreover, we follow the Breese et al. (1998) experimental design where the error is computed upon the same observed and test items for each user.

Following the *best-item* approach, we first dispatch a training set of 35 ratings and compute the predictions for the remaining 15 available ratings using the *lifestyle* and Pearson-based approaches. Then we develop the "success" table by comparing the performance of the two approaches. The class of each of the training examples is represented by the "success" or "failure" of the lifestyle approach compared to the Pearson-based one. The "success" table is utilized for the training of the Naïve Bayes algorithm, which is then consecutively queried for each of the 15 test-items and for each user in the sample. The output of each query is the selection of either the *lifestyle* or the Pearson-based approach as the most likely to be the best performing for the specific item given the target user's available ratings. For comparison purposes the "optimum *best-item"* is also included in Table 7 referring to an ideal algorithm which makes 100% accurate predictions concerning the best performing algorithm on the target item. In addition, we present the *best-item* performance when the Bayesian Networks algorithm (Mitchell 1997) is used as the meta-learning algorithm.

The above-results indicate that the *best-item* approach that utilizes the Naïve Bayes algorithm to predict the best performing algorithm on each item performs significantly better than the Pearson-based approach at high-density levels, though other learning algorithms may give better results approximating the performance of the optimum "best-item."

The proposed approach is a new hybridization technique where individual predictions are made independently from each other before they are combined. Its theoretical foundations are based on the stacking framework in machine learning research (Dzeroski and Zenko 2004). More specifically, it is a meta-learning approach that operates upon a set of base learning algorithms and a set of meta-features associated with their previous performance, which are represented by the "success" table. The best-item approach is characterized by its ability to accommodate any learning algorithm as well as by presenting item-level sensitivity without requiring the analysis of the item into features, as in CBF discussed next.

6.2 A content-based approach

In contrast to CF, CBF disregards all other users in the database and predicts future ratings based solely on a user's previous preference history (as expressed through ratings). In computational terms the most important advantage of CBF in comparison to CF is that it can produce recommendations even if no user has rated the target item, such as in the case when a new item is introduced in the database. Along these lines we

present a CBF approach designed for our application domain that may complement the CF-based approaches discussed so far when the above conditions occur.

In several domains it is rather difficult if feasible at all, to describe items by features as required by CBF. In particular in the domain under examination where advertised products are involved, it is a cumbersome task define low-level product features (e.g., price or color), which are universally applicable and meaningful to a wide range of products. It is not surprising that the majority of current work in adaptive and recommender systems which concerns product-oriented prediction refers mainly to specific product categories such as cars (Jameson et al. 1995), telephony devices (Ardissono and Goy 2000), books (Mooney and Roy 2000), restaurants (Pazzani 1999), and movies (Basu et al. 1998) in which the definition of common low-level features is feasible. In the proposed implementation of CBF, we examine the use of product subcategory (e.g., sports cars, family cars, insurance services, beers, soft drinks, and so on) as a single feature descriptive for all products that may associate previous preferences with rating predictions.

In our implementation of CBF, we select the Naïve Bayes algorithm, in which the user ratings represent five class labels. The algorithm is trained upon the available—for each user —{feature, class} pairs and when a product's feature is provided as input it predicts its class membership by computing the conditional probability $P(c|\text{feature}_1, \text{feature}_2,..,\text{feature}_n)$, where $c = 1, 2, 3, 4, 5$, and n represents the number of features. Naïve Bayes has been successfully used in content-based recommender systems in domains such as books and movies (Mooney and Roy 2000; Melville et al. 2002).

One of the important limitations of content-based algorithms is that they require a sufficient number of training examples in order to produce reliable predictions. For example, in cases where as many as 2, 5, or even 10 ratings are available for the target user, CF is clearly preferred over CBF due to the low levels of coverage (the items for which a prediction can be made by CBF). Thus, in the following experiment we evaluate the CBF performance on the "Given 50" protocol and compare with the CF results presented in the previous section.

The empirical findings demonstrate that CBF based on product subcategory is outperformed by both *lifestyle* and Pearson-based approaches (Table 8).

It must be noted that within specific product categories more features which are meaningful for certain groups of products would eventually improve the accuracy of prediction. Furthermore, it is possible to increase the performance of the CBF by employing regression techniques (Duda et al. 2000), which can directly predict numerical ratings instead of representing them as classes ignoring the linear scale.

The CBF performance depends on the amount of available ratings of the same subcategory with the target item. It is expected that if the target user has rated enough items belonging in the "same subcategory" with the target item then a more reliable CBF prediction can be made, since we have stronger indications concerning whether or not the user has liked advertised products of a specific product subcategory. Indeed, examining the subcategories of the training and test items, an association with the respective performance of CBF is revealed (Fig. 2).

Table 8 CBF performance for the selected features compared to CF approaches

CBF (Product subcategory)	Pearson-based	Lifestyle
0.9387	0.7528 $t = -4.636$; $p = 0.000$	0.7648 $t = -4.274$; $p = 0.000$

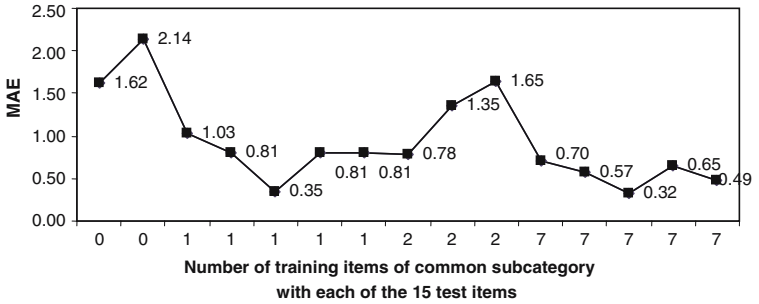

Fig. 2 CBF performance in relation to the subcategory occurrences in the training set

Table 9 A *best-item* combination of CBF and *lifestyle* producing significantly better results than any individual approach

	Content-lifestyle *Best-item*	Lifestyle	Pearson	CBF (subcategory)
MAE	0.6918	0.7647 $t = -6.922$ ($p < 0.0001$)	0.7528 $t = -5.267$ ($p < 0.0001$)	0.9387 $t = -5.767$ ($p < 0.0001$)

The highest errors concern the two items with zero occurrences of the same product subcategory in the training set. In this case, the algorithm outputs the most frequently observed class in the training set. Removing these two items, the effect on the averaged performance is significant but still CBF's MAE (0.7941) is higher than Pearson-based and *lifestyle* approaches (0.7369 and 0.7495, respectively). However, in the case where as many as seven items belong in the same subcategory with the test item then the CBF's performance is impressively improved (0.6497).

Along these findings, the two methods can be combined and exploit each other's advantages in order to provide more accurate prediction. Indeed, the combination of CBF and a collaborative filtering approach (e.g., the *lifestyle* one) under the *best-item* approach can give significantly improved performance compared to any of the individual approaches (Table 9).

The recommendation approaches discussed so far present complementary predictive performances with respect to the sparsity problem and therefore they may be aggregated into a single personalization strategy.

7 A personalization strategy

A personalization strategy (or, in other words a prediction strategy) "*consists of one or more prediction techniques and a set of rules that determine which technique*(*s*) *to use*" (van Setten 2002). The proposed strategy is extended beyond the boundaries of CF and incorporates CBF which under certain conditions can be complementary to the collaborative approaches.

The proposed strategy operates upon a set of "conditions" which rule the performance of the individual approaches. The decision process that will be described in the following sections relies on theoretical and empirical findings presented in this article concerning the performance and suitability of the various approaches.

7.1 Factors affecting CF and CBF

The major factors affecting the performance or restricting the use of CF or content-based algorithms directly define the conditions that guide the selection of the appropriate approach in the personalization strategy:

(a) **Condition 1 (C1)**: *the number of items rated by the target user* affects CF algorithms since the computation of similarities between the target and the remaining of the users is performed upon overlapping ratings. It also affects CBF: few rated items by the target user result into inaccurate predictions.
(b) **Condition 2 (C2)**: *the number of users "similar" to the target user, who rated the target item*, affects CF performance since the prediction on the target item is computed as the weighted average of the neighbors' ratings for the target item.
(c) **Condition 3 (C3)**: *the number of items "similar" to the target item rated by target user* affects CBF since the prediction on the target item is based upon the ratings provided by the target user on similar (to the target) items. CF is not affected by C3.
(d) **Condition 4 (C4)**: *sparsity*. The sparsity problem occurs when few data (ratings) are available for each user in the database and—similarly to the first condition—mainly affects the computation of similarities among users. In contrast to the first condition, sparsity refers to the entire database and can be measured by the sum of rated items by each user to the total number of items times the number of users. The overall sparsity level is an estimator for the number of available ratings for the users in the database on average and it is a strong indicator of the unreliability of the prediction. In conjunction with the first condition it specifies the reliability of the prediction based on CF. For example, in the Pearson-based approach, even if C1 is satisfied then at high-sparsity levels the possibility of computing similarities upon a sufficient amount of overlapping ratings is significantly reduced. Sparsity does not directly affect CBF because it ignores the ratings available for the remaining of the users.

From the above list of the main conditions, two special cases of conditions can be dispatched:

(e) **The new-user or cold start problem (CS)** is a special case of C1 that occurs when a new user is introduced into the system with an empty set of rated items. This results into the inability of both CF / CBF algorithms to produce recommendations.
(f) **The first-rater or new item problem (FR)** is a special case of C2 which occurs when a new item is introduced in the database for which no ratings are available. CF methods fail to produce a recommendation for the target user when none of his/her neighbors have rated the item. At the initiation of the system both the first-rater and cold-start problems occur.

Table 10 presents the different levels of effect of the conditions upon the approaches.
Collaborative filtering

(a) *Pearson-based*: it is highly affected by the number of ratings of the target user, the number of co-raters of the target item and the sparsity level. It fails to operate and produce predictions when a new user or a new item is introduced into the system.

Table 10 The factors' effect on the personalization approaches varying from failure to operate ("fail"), no-effect ("−"), low ("+"), medium ("++"), or high effect ("+++")

Basic method	Approach	Conditions					
		C1	C2	C3	C4	FR	CS
Collaborative Filtering	Pearson-based	+++	+++	−	+++	*Fail*	*Fail*
	Lifestyle	++	+++	−	++	*Fail*	++
	Integrated	+	+	−	+	++	*Fail*
	Best-item	++	+++	−	++	*Fail*	++
Segment-level (classification-based)	Center-based	++	+++	−	++	*Fail*	++
	Expert-based	−	−	−	−	−	−
Content-based	Product subcategory (CBF)	−	−	+++	−	−	+++

(b) *Lifestyle*: the *lifestyle* approach does not depend on the available user ratings to measure similarities and outperforms the Pearson-based when few ratings are available or the sparsity level is high. Thus, despite the fact that predictions become more accurate when more items are available, it is capable of producing recommendations based on neighbors' ratings even in the complete absence of rated items by the target user. It is affected by the number of co-raters of the target item and fails to produce recommendations in the complete absence of neighbors. It provides equivalent predictions to the Pearson-based approach at low sparsity levels.

(c) *Integrated*: it is on average the best performing approach (also covering the hybrid approach). Compared to the Pearson-based and *lifestyle* approaches upon which is developed, it benefits from the advantage of *lifestyle* (not highly affected by C1) and avoids the disadvantages of the Pearson-based (high effect of C2 and C4). Thus, in computational terms the effect of C1, C2, and C4 conditions is low.

(d) *Best-item*: it is useful at low-sparsity levels since it requires a sufficient number of ratings available in order to "learn" the best performing approach. As a meta-learning hybrid approach which does not predict ratings but algorithms, its performance is measured upon its ability to select the appropriate approach. Clearly the prediction accuracy concerning the target item relies on the selected approach (which in turn is affected by the current conditions).

Segment-level

(e) *Expert-based and center-based*. The expert-based is a special case of the segment-level approach. When static data are utilized as classificatory data then the approaches are not affected by the number of items rated by the users. The prediction in the center-based approach is formulated by aggregating the ratings of the target item available from the co-members of the lifestyle segment, thus being affected by the number of co-raters of the target item. However, in the expert-based, the classification of the users in the segments enables the direct assignment of predictions associated to the specific segment by a marketing expert (including the exploitation of large-scale psychographic surveys). The expert-based approach enables the prediction when either of the FR or CS conditions occurs. Furthermore, it is suitable for the initiating phase of the system

when both conditions occur and no other approach can make a prediction, since it does not depend on the number of ratings nor the density conditions in the database.

Content-based filtering

(f) *Product-subcategory CBF:* CBF based on product subcategory gives less accurate predictions than CF approaches but it does operate when very few or none of the remaining users have rated the target item. This is the only case in the proposed strategy for which CBF is considered as an alternative approach. The number of items rated by the target user does not have a direct effect on CBF since, even in the case that few ratings are available, the accuracy of prediction depends on whether those items are similar to the target item, in which case CBF can make a prediction.

7.2 Computation of factors' threshold values

The personalization strategy operates upon a query for prediction on the target item given the users' ratings, and compares the current values of the "factors" against a certain threshold. These factors' values can be computed as follows:

(a) C1(user) = number of ratings of the target user, which can be easily counted directly from the target user's ratings vector.
(b) C2 (user, database, item) = the number of similar users who have rated the target item. The arguments "user" and "database" are utilized to compute the similarities between the target and the remaining of the users. Then, the number of users who are similar to the target user and have rated the target item is returned.
(c) C3(user, item) = the number of items similar to the target item that have been rated by the target user. In the proposed implementation of CBF, this computation requires the matching of the target user's rated items to a product subcategory from the KB (that assigns products to subcategories). Then the number of rated items belonging in the specific product subcategory is returned.
(d) $\text{C4(database)} = \sum_{\text{user}_i \in \text{users}} \text{C1}(\text{user}_i) \big/ (\text{users} * \text{items})$. The sparsity score is easily computed from the database applying this formula, where $\text{C1}(\text{user}_i)$ refers to the number of rated items from user_i.

The thresholds for each case can be pre-computed upon the available data (and continuously refined) or by utilizing existing research results. The threshold values refer to the points that a significant change in the relative performance of the approaches is observed, suggesting a shift in the strategy in order to ensure that the best approach is selected. Thus, the threshold values can be derived from the performance curves comparing the competing approaches under the specified conditions. Indicative threshold values (applicable in our domain) that affect the decision process in the personalization strategy can be extracted from the empirical results presented in this paper as well as from other published research:

(a) C1_threshold = 5 ratings. In Sect. 5.2.1, we demonstrated that for up to five rated items by the target user, *lifestyle* approach performs significantly better than Pearson-based and is statistically equivalent to the *integrated* approach. Furthermore, for five available ratings both approaches present a stabilized performance, suggesting that the unreliability of prediction observed in the Given 2

protocol has been settled. After this point the performances are statistically equivalent and the assessment of additional conditions is required, since other approaches can be considered as more suitable.

(b) C2_threshold = 20 neighbors. Previous empirical findings from published research (Herlocker et al. 1999, 2002) suggest the adequate size of the neighborhood concerning the reliable performance of the Pearson-based approach.

(c) C3_threshold = 7 items. The empirical results in Sect. 6.2 indicate that for at least seven similar items in terms of product subcategory the content-based prediction is reliable and gives better results than CF approaches.

(d) $\text{C4\_threshold} = \begin{cases} \text{high} \approx 92\text{–}100\% \\ \text{medium} \approx 25\text{–}92\% \\ \text{low} \approx 0\text{–}25\% \end{cases}$. The findings in 5.4.1 provide an estimation of the sparsity intervals associated with variation in the performance of the *lifestyle*, *integrated,* and *best-item* approaches. Indeed, up to the 92% sparsity (five items rated on average in a total of 65 items) items, *lifestyle* can be selected, while as the sparsity levels drop, *integrated* or *best-item* should be selected (for 5–50 and 50–65 items, respectively).

7.3 Combining the approaches into a personalization strategy

The personalization strategy (Fig. 3) is represented by a decision tree where each node corresponds to a condition. Each condition is evaluated against the threshold value and depending on whether the condition is satisfied (denoted by "yes" in Fig. 3), the appropriate branch is followed, ending to an appropriate approach for the given condition values.

The decision concerning which of the major methods (CF / CBF) should be selected depends on the evaluation of condition C2. Indeed, if the number of neighbors who have rated the target item is sufficient then a CF approach should be selected (the left branch emanating from C2 = yes). It is worth noting that for traditional CF approaches

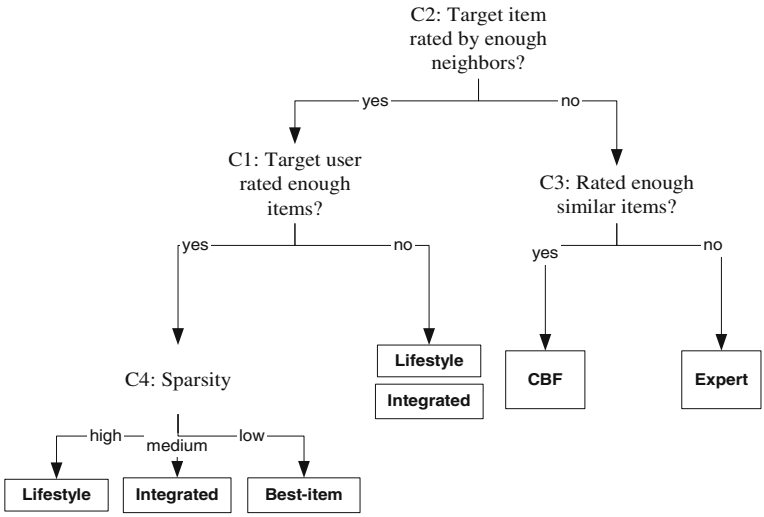

Fig. 3 The personalization strategy

such as the Pearson-based, C2 alone is not sufficient to justify its selection since it additionally requires enough ratings available for the target user in order to enable the measurement of similarities. On the contrary, lifestyle-based approaches can be selected either in the presence (C1 = yes) or in the absence (C1 = no) of a sufficient amount of rated items by the target user. If the target user has rated enough items (C1 = yes) then the sparsity level would determine which of the *lifestyle*, *integrated* or *best-item* is preferable. If the sparsity level is high, then *lifestyle* or *integrated* are appropriate techniques while *integrated* should be chosen if sparsity is medium. At low-sparsity levels the *best-item* approach presents the best performance. On the other hand, if C1 = no, then few ratings are available and the possible options include *lifestyle* and *integrated* since the *best-item* approach requires a sufficient amount of ratings in order to operate.

If C2 is not satisfied (C2 = no) then CF cannot make reliable predictions and CBF should be selected provided that the number of similar items rated by the target user is sufficient (C3 = yes), otherwise the expert-based approach is selected (C3 = no).

The strategy described incorporates the "new user" and "new item" conditions by applying zero threshold values for conditions C1 and C2, respectively (C1_threshold = 0, C2_threshold = 0).

An example of application of the above strategy is described in Table 11, where the objective is to find the appropriate recommendation approach for a target user and a target item in a given dataset consisting of n users and m items.

The proposed strategy can be simplified if we consider only the sparsity condition, which can be regarded as the most generic one in a recommender system's lifecycle (Fig. 4). This strategy suggests the *lifestyle* approach at the early stages of the recommender system's lifecycle, the *integrated* as more items become available and the *best-item* when the sparsity levels are low.

The computed sparsity value estimates the number of rated items to the total number of available items. It is reasonable to assume that after a period of system

Table 11 An application example of the personalization strategy

Action		Steps
1	Compute C2 (number of neighbors rated the target item)	• Compute similarities using an appropriate measure (e.g., Pearson) between the target and the remaining users • Select neighbors above a similarity threshold (e.g., threshold = 0) • Count the number of neighbors who have rated the target item (*assume* 20 *neighbors*)
		C2 = Yes
2	Compute C1 (number of items rated by the target user)	• Count the number of rated items by the target user (*assume* 10 *ratings available*)
		C1 = Yes
3	Compute C4 (database sparsity)	• Assume $n = 1{,}000$, $m = 500$, and total of available ratings = 5,000 • Sparsity = (1,000*500)/5,000 = 10% —*assume this percentage represents low-sparsity condition* (depends on the domain)
		C4 = low sparsity
Selected approach: *integrated* (or any other approach that deals with sparsity)		

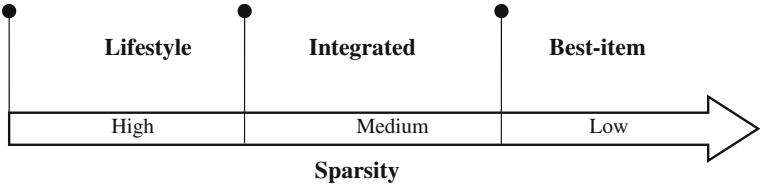

Fig. 4 A "linear" personalization strategy for collaborative filtering predictions

operations C1 and C2 will converge to meet the threshold values: the number of ratings for each user will increase as well as the number of neighbors of the target user (as more users are added into the system in particular for large databases). However, it is also apparent in real-world recommender systems that the same assumption may not be true for C4. Indeed, the ratio of rated items to the total number of items may remain low even after a long period of system usage due to the great amount of available items (e.g., book titles in Amazon can be tens of thousand) or the continuous introduction of new items (such as new products), or the increased unwillingness of the users to explicitly evaluate the items. Furthermore, the sparsity condition provides a quick and very generic assessment for a batch prediction concerning several users, rather than estimating all conditions described above for each one of the target user.

8 Summary and conclusions

In the present research different algorithmic approaches that utilize lifestyle data are proposed addressing inherent limitations of CF algorithms related to the sparsity of available ratings.

First, the proposed segment-level approach represents a direct implementation of target marketing methods based on lifestyle segmentation. It introduces a classification scheme based either on behavioral data (ratings) or on rating-independent user data in contrast to traditional classification methods based on extensive psychographic questionnaires. In the case of classification based on rating-independent data, the expert-based variation of the proposed approach addresses the cold-start problem, one of the most important drawbacks of CF approaches.

The lifestyle approach represents the fundamental construct toward the improvement of current personalization approaches. Since it requires no behavioral data to compute similarities among users and therefore it is applicable under the "new user" problem, in contrast to extant CF approaches which fail to make predictions. However, the most important contribution of the lifestyle approach is that it serves as the key constituent of the subsequent proposed approaches which further improve the predictive accuracy of CF algorithms in sparsity conditions.

The integrated approach is based upon and extends the hybrid approach. It presents its highest accuracy at low-to-medium sparsity levels. It efficiently manages the sparsity problem by substituting the original sparse user $\times$ item table by a dense table in which missing values are replaced by lifestyle-based predictions. Furthermore, it can accommodate any type of personalization approach that operates upon ratings and improve its performance through the reduction of the sparsity effect.

The best-item approach introduces the combination of lifestyle and Pearson-based ones but most importantly it proposes a new hybridization technique. Following Burke's (2002) taxonomy it can be classified as a "*switching*" hybrid since it switches

from one approach to another when specific conditions occur. In contrast to existing hybridization techniques, the best-item approach utilizes probabilistic learning algorithms trained upon the performance history of the rival approaches and predicts which of them will perform better for the target item. In this way, the proposed approach presents item-level sensitivity and avoids the interaction between the individual approaches, which may increase the final prediction error. The best-item approach presents superior performance to the Pearson-based as well as lifestyle approaches at low-sparsity level, complementing the performances of the lifestyle and integrated approaches. Furthermore, it presents high generalization abilities since it can accommodate any number of prediction approaches and increase their performance.

The proposed content-based approach based on product subcategory is in terms of predictive performance less accurate than CF methods, but when very few users have rated the target item, it demonstrates superior performance provided that the target user has rated a sufficient amount of items of the target item's subcategory. Thus it represents a suitable alternative approach when the above condition occurs.

Having defined the personalization framework through the development of several approaches that exploit the notion of lifestyle, a personalization strategy is proposed suggesting a suitable approach for a given condition. The personalization strategy provides a single framework incorporating the proposed approaches while it can be adjusted to accommodate any other approach since it is based upon well established criteria of the recommender systems theory.

Future research involves algorithmic improvements that can be further examined including the utilization of more descriptive product features (besides product subcategory) as well as the investigation of model-based algorithms and their effect on the predictive performance of the lifestyle-based approaches. In addition, the identification of suitable and domain independent lifestyle indicators that may improve the accuracy of the proposed approaches, is a direction for future research. The notion of lifestyle as a central construct of the consumer behavior model can be utilized to infer behaviors independently from the application domain. The extension of the empirical evaluation of the proposed approaches in larger samples as well as in domains such as Web-based product recommendation systems, or personalized services over mobile platforms, represent the most important future research avenue.

References

Aggarwal, C., Wolf, J., Kun-Lung, W., Yu, P. S.: Horting hatches an egg: a new graph-theoretic approach to collaborative filtering. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 201–212. San Diego, CA (1999)

Alspector, J., Koicz, A., Karunanithi, N.: Feature-based and clique-based user models for movie selection: a comparative study. User Model. User Adapt. Interact. **7**, 297–304 (1997)

Ardissono, L., Console, L., Torre, I.: An adaptive system for the personalized access to news. AI Commun. **14**(3), 129–147 (2001)

Ardissono, L., Goy, A.: Tailoring the interaction with users in web stores. User Model. User Adapt. Interact. **10**, 251–302 (2000)

Arndt, D., Gersten, W.: Data management in analytical customer relationship management. In: Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery, Workshop on Data Mining for Marketing Applications, pp. 25–38. Freiburg, Germany (2001)

Balabanovic, M.: An adaptive web page recommendation service. In: ACM First International Conference on Autonomous Agents, pp. 378–385. Marina del Ray, CA (1997)

Balabanovic, M., Shoham, Y.: Fab: content-based collaborative recommendation. Commun. ACM **40**(3), 66–72 (1997)
Basu, C., Hirsh, H., Cohen, W.: Recommendation as classification: using social and content-based information in recommendation. In: Proceedings of the 15th National Conference on Artificial Intelligence, pp. 714–720. Madison, WI (1998)
Beatty, S.F., Homer, P.M., Kahle, L.R.: Problems with VALS in international marketing research: an example from an application of the empirical mirror technique. Adv. Consum. Res. **15**, 375–380 (1998)
Belch, G.E., Belch, B.A.: Introduction to advertising and promotion: an integrated marketing communications perspective. Irwin, San Diego, CA (1995)
Billsus, D., Pazzani, M.: Learning collaborative information filters. In: Proceedings of the Fifteenth National Conference on Machine Learning, pp. 46–54. San Francisco, CA (1998)
Bozios, T., Lekakos, G., Skoularidou, V., Chorianopoulos, K.: Advance techniques for personalized advertising in a digital TV environment: the iMedia system. In: eBusiness and eWork Conference, Venice, Italy, pp. 1025–1031 (2001)
Breese, J.S., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the Fourteenth Annual Conference on Uncertainty in Artifical Intelligence, pp. 43–52. San Francisco, CA, (1998)
Brusilovsky, P.: Methods and techniques of adaptive hypermedia. User Model. User Adapt. Interact. **6**(2–3), 87–129 (2001)
Burke, R.: Hybrid recommender systems: survey and experiments. User Model. User Adapt. Interact. **12**, 331–370 (2002)
Cawley, G.C., Talbot, N.L.C.: Efficient leave-one-out cross-validation of kernel Fisher discriminant classifiers. Pattern Recognit. **36**(1), 2585–2592 (2003)
Chaney, D.: Lifestyles. Routledge, London (1996)
Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., Sartin, M.: Combining content-based and collaborative filters in an online newspaper. In: ACM SIGIR Workshop on Recommender Systems, Berkeley, CA (1999)
Domingos, P., Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one-loss. Mach. Learn. **29**, 103–130 (1997)
Domshlak, C., Joachims, T.: Efficient and non-parametric reasoning over user preferences. User Model. User Adapt. Interact. (this issue) (2006)
Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley, New York (2000)
Dzeroski, S., Zenko, B.: Is combining classifiers with stacking better than selecting the best one? Mach. Learn. **54**, 255–273 (2004)
Goldberg, D., Nichols, D., Oki, B., Borchers, A.: Using collaborative filtering to weave and information tapestry. Commun. ACM **35**(12), 61–70 (1992)
Goldberg, K., Roeder, T., Guptra, D., Perkins, C.: Eigentaste: a constant-time collaborative filtering algorithm. Inform. Retrieval **4**(2), 133–151 (2001)
Good, N., Schafer, J.B., Konstan, J., Borchers, A., Sarwar, B., Herlocker, J., et al.: Combining collaborative filtering with personal agents for better recommendations. In: Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-99), Menlo Park, CA, pp. 439–446 (1999)
Gunter, B., Furnham, A.: Consumer Profiles: An Introduction to Psychographics. Routledge, London (1992)
Hair, J.F., Anderson, R.E., Tatham, R.L., Black, W.C.: Multivariate Data Analysis. Prentice Hall London (1998)
Harter, S.P.: Variations in relevance assessments and the measurement of retrieval effectiveness. J. Am. Soc. Inform. Sci. **47**(1), 37–49 (1996)
Hawkins, I., Best, R.J., Coney, K.A.: Consumer Behavior: Building Marketing Strategy. Irwin/McGraw-Hill New York (1998)
Herlocker, J., Konstan, J.: Content-Independent task-focused recommendation. IEEE Internet Comput. **5**, 40–47 (2001)
Herlocker, J., Konstan, J., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. In: Proceedings of the Twenty-second International Conference on Research and Development in Information Retrieval (SIGIR'99), New York, pp. 230–237 (1999)
Herlocker, J., Konstan, J., Riedl, J.: An empirical analysis of design choices in neighborhood-base collaborative filtering algorithms. Inform. Retrieval **5**, 287–310 (2002)
Herlocker, J., Konstan, J., Terveen, L.G., Riedl, J.: Evaluating collaborative filtering recommender systems. ACM Trans. Inform. Syst. **22**(1), 5–53 (2004)

Hill, W., Stead, L., Rosenstein, M., Furnas, G.: Recommending and evaluating choices in a virtual community of use. In: ACM CHI'95 Conference on Human Factors in Computing Systems, Denver, Colorado, 194–201 (1995)
Hoffman, T., Puzicha, J.: Latent class models for collaborative filtering. In: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, Stocholm, Sweden, 688–693 (1999)
Hofmann, T.: Latent semantic models for collaborative filtering. ACM Trans. Inform. Syst. **22**(1), 89–115 (2004)
Hollink, M., van Someren, M., Wielinga, B.: Discovering stages in web navigation for problem-oriented navigation support. User Model. User Adapt. Interact. (this issue) (2006)
Jameson, A., Schafer, R., Simons, J., Weis, T.: Adaptive provision of evaluation-oriented information: tasks and techniques. In: Proceedings of the Fourteenth international joint conference on Artificial Intelligence, Montreal, Canada, 1886–1893 (1995)
Kara, A., Kaynak, E.: Markets of a single customer: exploiting conceptual developments in market segmentation. Eur. J. Mark. **31**(11/12), 873–895 (1997)
Karypis, G.: Evaluation of item-based top-N recommendation algorithms. In: CIKM 2001, Atlanta, GA, pp. 247–254 (2001)
Kobsa, A., Koenemann, J., Pohl, W.: Personalized hypermedia presentation techniques for improving online customer relationships. Knowl. Eng. Rev. **16**(2), 111–155 (2001)
Kotler, P.: Marketing Management. Prentice-Hall Englewood Cliffs, NJ (1994)
Krulwich, B. Lifestyle finder: intelligent user profiling using large-scale demographic data. AI Mag. 37–45 (1997)
Lekakos, G., Giaglis, G.: Personalization of advertisements in the digital TV context. In: Shen C., Magoulas, G. (eds.) Adaptable and Adaptive Hypermedia Systems, pp. 264–283. Idea Group PA (2005)
Lekakos, G., Giaglis, G.: Improving the prediction accuracy of recommendation algorithms: approaches anchored on human factors. Interact. Comput. **18**(3), 410–431 (2006)
Lekakos, G., Giaglis, G.M.: A lifestyle-based approach for delivering personalized advertisements in digital interactive television. J. Comput. Mediated Commun. **9**(2) (2004)
Malone, T.W., Grant, K.R., Turbak, F.A., Brobst, S.A.: Intelligent information systems. Commun. ACM **30**(5), 390–402 (1987)
Maybury, M.: PersonalCasting. In: Workshop on Personalization in Future TV, Sonthofen, Germany, retrieved on 5/6/2005 from www.di.unito.it/∼liliana/UM2001 (2001)
Melville, P., Mooney, R.J., Nagarajan, R.: Content-boosted collaborative filtering for improved recommendations. In: Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI), pp. 187–192. Edmonton, Canada (2002)
Milenkovic, M.: Delivering interactive services via a digital TV infrastructure. IEEE Multimedia **5**(4), 34–43 (1998)
Mooney, R. J., Roy, L.: Content-based book recommending using learning for text categorization. In: Proceedings of the Fifth ACM Conference in Digital Libraries, San Antonio, TX, pp. 195–204 (2000)
Moore, A.W., Lee, M.S.: Efficient algorithms for minimizing cross validation error. In: Proceedings of the 11th International Conference on Machine Learning, San Francisco, CA, pp. 190–198 (1994)
Mowen, J.C., Minor, M.: Consumer Behavior, 5th edn. Prentice-Hall Upper Saddle River, New Jersey (1998)
O'Mahony, M., Hurley, N., Silvestre, G.: Promoting recommendations: an attack on collaborative filtering. In: Proceedings of the Thirteenth International Conference on Database and Expert Systems Applications, Aix-en-Provence, France, pp. 494–503 (2002)
Pazzani, M.: A framework for collaborative, content-based and demographic filtering. Artif. Intell. Rev. **13**(5/6), 393–408 (1999)
Pazzani, M., Billsus, D.: Learning and revising user profiles: the identification of interesting web sites. Mach. Learn. **27**, 313–331 (1997)
Peltier, J.W., Schibrowsky, J.A., Schultz, D.E., Davis, J.: Interactive psychographics: cross-selling in the banking industry. J. Advert. Res. **42**(2), 7–22 (2002)
Pennock, D.M., Horvitz, E., Lawrence, S., Giles, C.L.: Collaborative filtering by personality diagnosis: a hybrid memory and model-based approach. In: Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence, San Francisco, CA, pp. 473–480 (2000)
Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: GroupLens: an open architecture for collaborative filtering of NetNews. In: ACM Conference on Computer Supported Cooperative Work, pp. 175–186. Chapel Hill, NC (1994)
Resnick, P., Varian, H.R.: Recommender systems. Commun. ACM **40**(3), 56–58 (1997)
Rich, E.: User modeling via stereotypes. Cognit. Sci. **3**, 329–354 (1979)

Rich, E.: Users are individuals: individualizing user models. Int. J. Man-Mach. Stud. **18**, 199–214 (1983)
Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Analysis of recommendation algorithms for e-commerce. In: ACM e-commerce conference, Minneapolis, Minnesota, pp. 158–167 (2000)
Sarwar, B.M., Karypis, G., Konstan, J., Riedl, J.: Item based collaborative filtering recommendation algorithms. In: Proceedings of the Tenth International World Wide Web Conference, pp. 285–295 (2001)
Sarwar, B.M., Konstan, J., Borchers, A., Herlocker, J., Miller, B., Riedl, J.: Using Filtering Agents to Improve Prediction Quality in the Grouplens Research Collaborative Filtering System, CSCW, Seattle, WA, pp. 345–354 (1998)
Schafer, J.B., Konstan, J.A., Riedl, J.: Electronic-commerce recommender systems. J. Data Mining Knowl. Discov. **5**(1), 115–152 (2001)
Schwab, I., Pohl, W., Koychev, I.: Learning to recommend from positive evidence. In: Proceedings of the Intelligent User Interfaces, New Orleans, LA, pp. 241–247 (2000)
Shardanand, U., Maes, P.: Social information filtering: algorithms for automating "Word of Mouth" In: ACM CHI'95 Conference on Human Factors in Computing Systems, Denver, Colorado, pp. 210–217 (1995)
Smyth, B., Cotter, P.: A personalized television listings service. Commun. ACM **43**(8), 107–111 (2000)
Ungar, L., Foster, D.: Clustering methods for collaborative filtering. In: Workshop on Recommendation Systems at the Fifteenth National Conference on Artifical Intelligence, Madison, Wisconsin, pp. 112–128 (1998)
van Setten, M., Veenstra, M., Nijholt, A.: Prediction strategies: combining prediction techniques to optimize personalization. In: AH '2002 Workshop on personalization in Future TV, Malaga, Spain, pp. 29–38 (2002)
Vyncke, P.: Lifestyle segmentation: from attitudes, interests and opinions, to values, aesthetic styles, life visions and media preferences. Eur. J. Commun. **17**(4), 445–464 (2002)
Webb, G., Pazzani, M., Billsus, D.: Machine learning for user modeling. User Model. User Adapt. Interact. **11**, 19–29 (2001)

Authors' vitae

Dr. George Lekakos Athens University of Economics and Business, Department of Management Science and Technology, 47 Evelpidon Str., 11362 Athens, Greece. Dr. Lekakos is an adjunct Lecturer at the Department of Management Science and Technology, Athens University of Economics and Business, Athens, Greece and a Visiting Lecturer at the Department of Computer Science, University of Cyprus. He is the director of the Digital Interactive Media (DIM) research team of the ELTRUN research group within the Athens University of Economics and Business. He holds a B.Sc. in Mathematics from the University of Thessaloniki, Greece, an M.Sc. in Formal Methods in Software Engineering from the Queen Mary and Westfield College, University of London, UK, and a Ph.D. from the Department of Management Science and Technology, Athens University of Economics and Business, Greece. Dr. Lekakos has worked in the area of personalized and adaptive systems, human–computer interaction, and machine learning with emphasis on the development of personalization algorithms. He has published more than thirty papers in international journals and conferences, and he is the co-editor of books and conference proceedings.

Dr. George Giaglis Athens University of Economics and Business, Department of Management Science and Technology, 47 Evelpidon Str., 11362 Athens, Greece. Dr. Giaglis is an Associate Professor of eBusiness at the Department of Management Science and Technology of the Athens University of Economics and Business, Greece. He has also held full-time academic posts in Brunel University (UK) and the University of the Aegean (Greece), while he has been a visiting professor in universities such as the University of London, Nottingham Trent University, and Henley Management College. His main teaching and research interests lie in the areas of eBusiness (emphasising on mobile and wireless applications and services), technology-enabled business process redesign, business process modeling and re-engineering, information management, and information systems evaluation. He has published more than 50 research articles in leading journals and international conferences. He is a member of the editorial board of the International Journal of Mobile Communications and the Logistics Information Management Journal. He is the Director of the ELTRUN Wireless Research Group (ELTRUN/WRC) hosted by the Athens University of Economics and Business, pursuing research in innovative mBusiness models, assessment of new mobile/wireless technologies, applications and services development, and others.