User Model User-Adap Inter (2012) 22:151–175
DOI 10.1007/s11257-011-9111-y

ORIGINAL PAPER

Discovery of Web user communities and their role in personalization

Georgios Paliouras

Received: 2 October 2009 / Accepted in revised form: 31 December 2010 /
Published online: 10 March 2012

Abstract One of the major innovations in personalization in the last 20 years was the injection of social knowledge into the model of the user. The user is not considered an isolated individual any more, but a member of one or more communities. User communities have been facilitated by the striking advancements of electronic communications and in particular the penetration of the Web into people's everyday routine. Communities arise in a number of different ways. Social networking tools typically allow users to proactively connect to each other. Alternatively, data mining tools discover communities of connected Web sites or communities of Web users. In this article, we focus on the latter type of community, which is commonly mined from logs of users' activity on the Web. We recall how this process has been used to model the users' interests and personalize Web applications. Collaborative filtering and recommendation are the most widely used forms of community-driven personalization. However, we examine a range of other interesting alternatives that are worth investigating further. This effort leads us naturally to the recent developments on the Web and particularly the advent of the social Web. We explain how this development draws together the different viewpoints on Web communities and introduces new opportunities for community-based personalization. In particular, we propose the concept of active user community and show how this relates to recent efforts on mining social networks and social media.

Keywords User communities · Web mining · Web personalization · Web communities · Social networks

G. Paliouras (✉)
Institute of Informatics and Telecommunications, National Centre for Scientific Research "Demokritos", Patr. Grigoriou & Neapoleos str., Ag. Paraskevi, 15310 Attiki, Greece
e-mail: paliourg@iit.demokritos.gr

1 Introduction

Paradoxically, one of the most successful and large-scale uses of personalization technology, collaborative filtering and recommendation, does not focus on *what* an individual likes or dislikes, but rather on *who* this individual is related to. Beyond the many technical advantages of this approach, there is an important social advancement that has happened hand-in-hand with it. That is the emergence of the *social Web* or *Web 2.0*, which made Web users active participants, generating their own content and forming on-line social networks. In this process, the focus moved from the individual to the communities in which he or she belongs.

User modeling aims to understand the needs and interests of the user and produce information systems that can adapt their behavior to these personal requirements. In the beginning of the 1990s, as the use of the Web started spreading beyond computer experts, user modeling researchers observed two new opportunities for personalization:

- The fact that the usage of Web-based information systems was easily recorded, either at the server or the client side. This has initiated a new breed of data-driven Web personalization methods and systems (e.g. Joachims et al. 1997; Pazzani et al. 1996) that were soon to become mainstream.
- The fact that important knowledge can be discovered about an individual, based on the communities in which he or she belongs (Hill et al. 1995; Orwant 1995; Shardanand and Maes 1995).

Soon after that, it became clear that the combination of the two was the natural way forward and could have a multiplying effect on our ability to understand the user and personalize Web-based systems. As a result, data-driven methods, adopting statistical and machine learning approaches, were developed to discover communities of users (Cooley et al. 1997; Orwant 1995). Such methods, in turn, supported new powerful ways of personalization, such as collaborative filtering and recommendation (Hill et al. 1995; Konstan et al. 1997; Konstan and Riedl 2012; Shardanand and Maes 1995).

The role of Web user communities in user modeling and personalization is the focus of this article. A full account and survey of related work is beyond its scope and the interested reader is referred to earlier efforts (e.g. Mobasher 2007; Pierrakos et al. 2003; Su and Khoshgoftaar 2009). Instead, this article attempts to clarify and explain the different uses of the term *Web community* (Sect. 2), before moving on to present how data mining has been used to identify user communities (Sect. 3) and support personalization (Sect. 4). Following this, the different types of Web community are revisited, in the light of the social Web, in order to sketch a new research opportunity for user modeling, namely the discovery of communities of *active Web users* (Sect. 5). Finally, this article concludes with a discussion of the more general implications of discovering communities in the social Web (Sect. 6).

2 Web communities

Based on its state of development in the early 1990s, the Web was viewed either as a networked information source or as an online marketplace of networked services, in

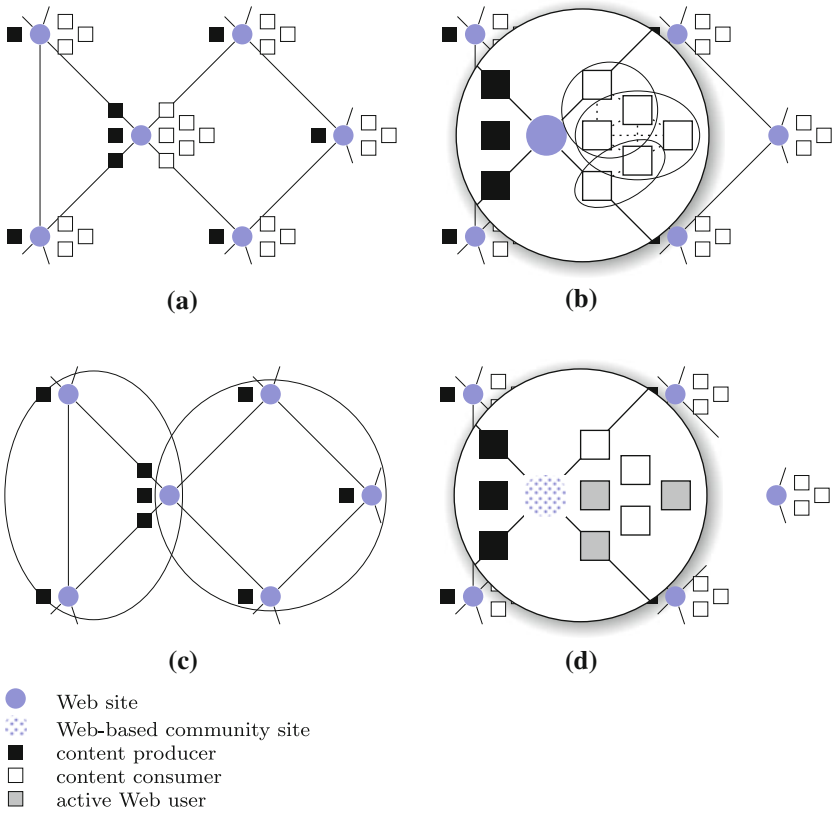

Fig. 1 A traditional view of the Web and the three types of community. **a** A sketch of the old Web, **b** Web usage communities within a site, **c** Web communities on the Web graph, and **d** Web-based community, as a single site

its infancy. Either way, there was a clear separation between producers and consumers of content on the Web. Figure 1a provides a simplified sketch of this situation. At each node s_i of the network S, there is a group of producers UP_i and a group of consumers UC_i of the content of the corresponding Web site. Although the two groups are not necessarily disjoint, they are treated separately, as the production process is not considered part of the usage of a Web node. Furthermore, the group of consumers is assumed to be considerably larger than that of producers, i.e. $|UC_i| >> |UP_i|$.

User modeling research has focused on the consumers UC_i of this set-up. Among the various reasons for this, were (a) the fact that they were many and of a more variable background, (b) the provision of technology for logging their activity at the servers of a Web site and (c) their role as customers of online shops, which aimed to increase their customer base. On this basis, interest has grown in discovering *Web user communities*, by analyzing the usage data logged at the Web servers. In more traditional economics terminology, Web user communities provided a segmentation of the consumers UC_i of s_i (see Fig. 1b). This segmentation was done in various ways, ranging from simple statistical analysis of demographic groups, often called

user stereotypes (Rich 1979), to the dynamic discovery of associations between users, based on the analysis of usage data by machine learning methods.

Definition 1 (*User Community*) Each segment $UC_{ij} \subset UC_i$ of consumers, commonly produced by the analysis of usage data of Web site s_i.

Although producers of Web content were not of obvious interest to personalization research, the content that they produced, i.e. the Web sites, and the way in which it was interconnected to provide the network of the Web, was of great interest to meta-services, such as search engines. As a result, a parallel strand of work was developed, which is still very active, and aims at discovering patterns in Web structure, often called *Web structure mining* (Kosala and Blockeel 2000). The end product of this process was typically a segmentation of the Web graph S into *Web communities* $S_j \subset S$ (Flake et al. 2000; Kumar et al. 1999) (see Fig. 1c). These communities can be linked to Web users only through the production process of Web sites. In the extreme case that a single user up_i is associated with the production of each s_i, a Web community is associated with a community of producers (Adamic and Adar 2003). The larger and more complex the Web site, the looser this link becomes. As a result, despite the use of similar statistical and machine learning approaches, Web communities have always been clearly separated from user communities, as studied in user modeling.

Definition 2 (*Web Community*) Each segment $S_j \subset S$ of the Web graph, produced by the analysis of Web graph structure.

The term Web community, more accurately *Web-based community*, was also used in a different area of research in information systems. Namely systems that support the formation or strengthening of real-life communities. These were typically either local communities (e.g. within a University campus) or interest-driven ones (e.g. professional associations). Early work in this field focused on *community networks* (Schuler 1994) and *virtual communities* (Rheingold 1993), which then evolved into *Web community portals* (Staab et al. 2000) and finally into *Web-based communities* (Bouras et al. 2005). In the context of Fig. 1a, typically Web-based communities correspond to special nodes s_j of the Web graph S, which serve as a meeting point for a community of users. Thus, all the users UC_j of the node explicitly join and make up the community. What makes this type of community particularly interesting is the fact that users start producing content for the community (see Fig. 1d). In other words, they move from being passive consumers of content UC_i to becoming active users UA_i of that particular Web site. This idea, aided by corresponding technological advances, later developed in what we now call the social Web. We will return to this in Sect. 5, where we will re-examine the three separate notions of community presented here.

Definition 3 (*Web-based Community*) A specific node $s_j \in S$ of the Web graph, serving as a meeting point for a community of users.

3 Discovering user communities from data

Having explained the differences from *Web communities*, i.e. segments of the Web graph, and *Web-based communities*, i.e. Web sites that support real-life communities,

we now focus on *Web user communities* and how these can be discovered in Web usage data. In the rest of this section, we will look at the usage data that are typically collected on the Web and recall how data mining approaches have been used to discover user communities from these data.

3.1 Usage data

The main assumption in the discovery of user communities is that the users do not have to explicitly register for the communities. As a result, community discovery has to be based on information provided by the users that may imply commonalities and associations among them. Typically, this information is taken to imply common interest of users in the products or services provided by a particular Web site. However, other types of association, such as level of expertise and collaboration in Web-based educational systems, have also been studied (Gaudioso and Boticario 2003). For the sake of simplicity, in the rest of this section we will focus on the interest of user uc_{il} in a particular *item*, i.e. product or service, t_{ik} of Web site s_i.

There are several observations that help us to infer the interest of a user in a particular item. Among them, the most common ones are:

- The selection of the item for viewing.
- The purchase of the item.
- The explicit rating of the item.

Typically this information is recorded in the databases of the Web site. The simplest form of such a database is the log of the Web server, which records all visits (hits) to the Web pages of the site. Table 1 provides an extract of a Web server log, showing, among other things, the IP address of the visitor, the time stamp of the visit and the Web page that the user visited.

Extended versions of server logs can record other information, such as the id of the user, if the user has logged in, or the referring Web page that the user was viewing before the hit. Such information can be particularly useful for making a more

Table 1 An extract from a simple Web server log file

127.1.2.2	[10/Sep/2010:21:15:05]	"GET /index.html HTTP/1.1"	200	1043
127.1.2.2	[10/Sep/2010:21:15:06]	"GET /main.html HTTP/1.1"	200	954
127.1.2.2	[10/Sep/2010:21:15:07]	"GET /books.html HTTP/1.1"	200	837
127.1.2.2	[10/Sep/2010:21:15:08]	"GET /books/p13.html HTTP/1.1"	200	568
204.0.0.1	[10/Sep/2010:21:15:10]	"GET /index.html HTTP/1.1"	200	1043
204.0.0.1	[10/Sep/2010:21:15:11]	"GET /cars.html HTTP/1.1"	200	1235
127.1.2.2	[10/Sep/2010:21:15:12]	"GET /books/p14.html HTTP/1.1"	200	2037
204.0.0.1	[10/Sep/2010:21:15:12]	"GET /cars/p38.html HTTP/1.1"	200	8923
204.0.0.1	[10/Sep/2010:21:15:15]	"GET /cars/p97.html HTTP/1.1"	200	9478
127.1.2.2	[10/Sep/2010:21:15:17]	"GET /books/p15.html HTTP/1.1"	200	4056
204.0.0.1	[10/Sep/2010:21:15:20]	"GET /extra/p29.html HTTP/1.1"	200	3459

accurate inference of the user's interest in a particular item (Wu et al. 1998). Even the identification of the user who visited a Web page is not straightforward if the user id is missing. Several alternative approaches, such as cookies and Javascript, have been used for identification, especially when visitors are not registered to the Web site. However, some of these methods are treated with suspicion by the users, as they constitute potential threats to their privacy (Cooley et al. 1999).

Usage data are pre-processed, in order to obtain a more accurate estimate of the user's interest. Examples of such processing are the removal of noise from the data, e.g. visits of robots to the site, and the association of Web pages with the corresponding items or item categories. One pre-processing step, that is of particular importance, is the identification of *user sessions*. A user session is a chronologically-ordered sequence of page hits that can be attributed to the same user and has a particular beginning and end. For instance, in Table 1, the sequence of pages visited by the IP address "127.1.2.2", constitutes one session, starting with page "/index.html" and ending with "/books/p15.html". In other words, the session captures the complete visit sequence of a particular user at a particular point in time.

User sessions provide associations among the items, i.e. items that are visited within the same session, e.g. "p13" and "p15" in the above example, can be considered to be associated. This is especially important when user identification is not possible and therefore a longer history of item selection by a user cannot be established. The identification of sessions in usage data is also not free of technical difficulties and a number of methods, such as a time-out period of 30 min between hits, have been proposed to overcome them (Srivastava et al. 2000). The longer a user session is, the more information it provides about the preferences of the user. However, the majority of user sessions are typically very short and can thus not provide sufficient information in isolation. For this reason, the aggregation of information from larger sets of user sessions, either through user identification or through the discovery of user communities, is essential.

The history of interaction of a user uc_{il} with s_i comprises a simple user model, often called *user profile*. Typically, such a user profile maintains numerical preference values of the user about the items provided on the Web site. Depending on the kind of user input that is recorded, these values may correspond, among others, to:

- Frequency of selection of the item.
- Viewing time of the item.
- Frequency of purchase of the item.
- Explicit rating of the item.

The set of user profiles for the consumers UC_i of the items T_i of s_i is therefore defined by an *interest* function $f_i : UC_i \times T_i \rightarrow \mathbb{R}$. The values of this function are stored in a user-item matrix, such as the one shown in Table 2a. Moving a step closer to personalization, the numeric values are often mapped onto binary ones, through a thresholding process that determines when an item is of interest or not to a user. Table 2b binarizes the profiles of Table 2a, using an interest threshold of 0.5, i.e. interest values below the threshold are interpreted as lack of interest of the user in the particular item. Furthermore, as is common in user modeling, the length of the user interaction history that is used for constructing the user profile may vary. In the simplest case of a very

Table 2 User-item matrices recording user interest

	$p13$	$p14$	$p15$	$p29$	$p38$
(a) Numeric user profiles					
u1	1.0	0.6	0.8	0.0	0.0
u2	0.9	0.3	0.0	0.8	0.0
u3	0.0	0.0	0.0	0.8	0.9
u4	0.6	0.0	0.0	1.0	0.0
u5	0.1	0.8	1.0	0.0	0.0
(b) Binary profiles (threshold 0.5)					
u1	1	1	1	0	0
u2	1	0	0	1	0
u3	0	0	0	1	1
u4	1	0	0	1	0
u5	0	1	1	0	0

short-term model, only the values of a single user session are taken into account. In this special case, if the user id is unknown, session profiles are constructed.

3.2 Identifying user relations

In order to identify communities, one needs to measure the degree of similarity between users, in terms of their expressed interest about items. This can be achieved by defining a similarity function $R_i : \mathbf{UC}_i \times \mathbf{UC}_i \rightarrow \mathbb{R}$, where $\mathbf{UC}_i$ is the set of user profile vectors for UC_i in the multidimensional item space (see Fig. 2a). Common choices are the cosine similarity and Pearson correlation, shown in (1) and (2) respectively. These measures give a high similarity value when the profiles of the two users are similar, e.g. the users have visited the same Web pages. A variety of other similar measures have been proposed in the literature (e.g. Herlocker et al. 1999), some being more suitable to binary and other to numerical profiles.

$$R_i(\mathbf{uc}_{il}, \mathbf{uc}_{im}) = \frac{\sum_{k=1}^{T_i} uc_{ilk} \times uc_{imk}}{\sqrt{\sum_{k=1}^{T_i} uc_{ilk}^2} \times \sqrt{\sum_{k=1}^{T_i} uc_{imk}^2}} \tag{1}$$

$$R_i(\mathbf{uc}_{il}, \mathbf{uc}_{im}) = \frac{\sum_{k=1}^{T_i} (uc_{ilk} - \overline{uc_{il}}) \times (uc_{imk} - \overline{uc_{im}})}{\sqrt{\sum_{k=1}^{T_i} (uc_{ilk} - \overline{uc_{il}})^2} \times \sqrt{\sum_{k=1}^{T_i} (uc_{imk} - \overline{uc_{im}})^2}} \tag{2}$$

Based on the similarity of the users, measured by R_i, one can construct a weighted graph $UG_i = (UC_i, UE_i, R_i)$, the vertices of which represent the users, and the edges UE_i, weighted by R_i, denote the degree of similarity among the users. Figure 2b provides an example of such a graph. The user graph can be used, instead of the item space, for community discovery. Furthermore, it is commonly transformed into an unweighted graph, by introducing a similarity threshold, below which users are

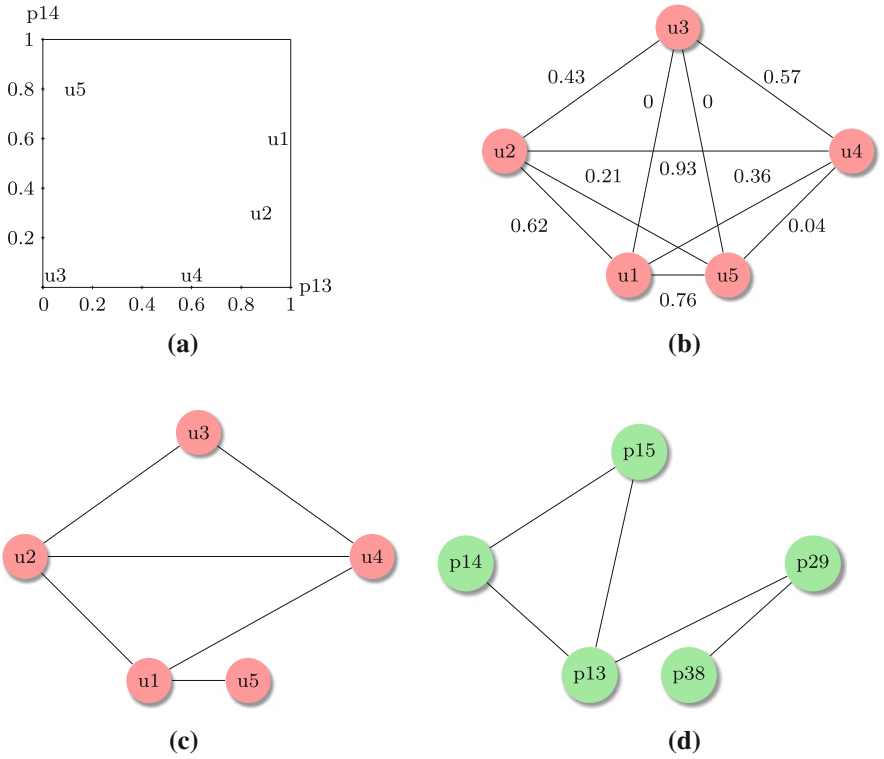

Fig. 2 Measuring similarity among user profiles and constructing the user and item graphs. **a** Profiles of the users in Table 2a, **b** Weighted graph of users in Table 2a, **c** Unweighted user graph (threshold 0.3), and **d** Unweighted graph of items in Table 2a

assumed to not be related. The choice of this threshold is another important decision towards personalization. Using the similarity threshold, the graph of Fig. 2b can be transformed into the graph of Fig. 2c. An important effect of this process is the reduction of the connectivity of the graph, which facilitates the use of various graph-theoretic discovery methods.

Instead or in addition to the user graph, personalization is often based on a dual representation: the *item graph*. The item graph is constructed in a similar manner to the user graph, by simply transposing the user-item matrix. Item graphs are particularly important when users cannot be identified and similarity is calculated on the basis of sessions. For instance, item similarity can be calculated as the frequency of co-occurrence of two items in the session profiles, when user profiles are not available.

By analogy to user graphs, the item similarity function is defined on the item set T_i, the item profiles are represented as vectors in the user space and the item graph $TG_i = (T_i, TE_i, R_i)$ is constructed, where R_i is now defined as $R_i : \mathbf{T}_i \times \mathbf{T}_i \rightarrow \mathbb{R}$, and $\mathbf{T}_i$ are the profile vectors of T_i. Figure 2d shows an example of an unweighted item graph. Item graphs capture similar information to user graphs, i.e. the associations between items are based on their usage. This is different to content-based approaches,

which are common in information retrieval (Baeza-Yates and Ribeiro-Neto 1999) and measure the similarity among items on the basis of characteristic features, i.e. the content and structure of the items themselves.

In the rest of this section, we will refer to communities discovered on both the user and the item graphs. Clearly, the communities constructed on item graphs are not segments of the set of users UC_i, as defined in Sect. 2. However, they can be considered to capture the preferences of latent communities of users and as such can be used for personalization (Mobasher et al. 2000; Paliouras et al. 2000). We will henceforth refer to them as *community models*, in order to distinguish them from user communities.

Definition 4 (*Community Model*) A vector $\mathbf{cm}_{ij}$ in the item space, similar to the models of individual users, which however represents the preferences of the corresponding community of users UC_{ij}.

A common special case is when the item space is binary, i.e. an item either exists or not in a model. In this case, the model is defined as a set of items, i.e. $\mathbf{cm}_{ij} = T_{ij} \subset T_i$. For instance, in the graph of Fig. 2d, the set of pages {"p13","p14","p15"} are closely related, i.e. they appear often together in user sessions, and thus may model the preferences of a community. In this sense, they form a community model. Community models are essential for personalization, even when communities of users are explicitly discovered from the usage data. This is because they reveal the preferences of the communities, which is actionable knowledge for a personalization system.

3.3 Mining for user communities

Having measured the similarity R_i among the users UC_i of a site s_i, communities are defined simply as clusters of similar users. Therefore, generic clustering methods have been used, in order to discover communities in usage data. The most widely researched and used type of clustering is *hard partitioning*, which aims to separate a set of objects into cohesive and well-separated subsets. The objects in our case are the users, represented by their profiles and the subsets are the communities, i.e. $UC_i = \bigcup_{j=1..J} UC_{ij}$. Hard partitioning methods aim to identify regions of the item space (see Fig. 2a) that are densely populated and isolated from each other. In other words, they maximize $R_i(\mathbf{uc}_{il}, \mathbf{uc}_{im})$, $uc_{il}, uc_{im} \in UC_{ij}$, within each community and minimize $R_i(\mathbf{uc}_{il}, \mathbf{uc}_{im})$, $uc_{il} \in UC_{ij}, uc_{im} \in UC_{ih}$, $j \neq h$, among communities. Figure 3a illustrates the clustering process. Early research in community discovery (e.g. Orwant 1995; Yan et al. 1996) adopted such hard partitioning approaches, including among others graph-theoretic partitioning of the user or item graphs (e.g. Kohrs and Mérialdo 1999; O'Connor and Herlocker 1999).

One of the problems with hard partitioning methods is that they assume each user to belong in one and only one community. This is problematic because users have diverse interests and may well belong in different communities. As a result, soon after these early efforts, approaches that allow the discovery of overlapping communities appeared. These approaches were typically based on common graph analysis methods, such as the identification of cliques and connected components. Figure 3b shows how the graph of Fig. 2d can be separated into overlapping cliques. Such methods

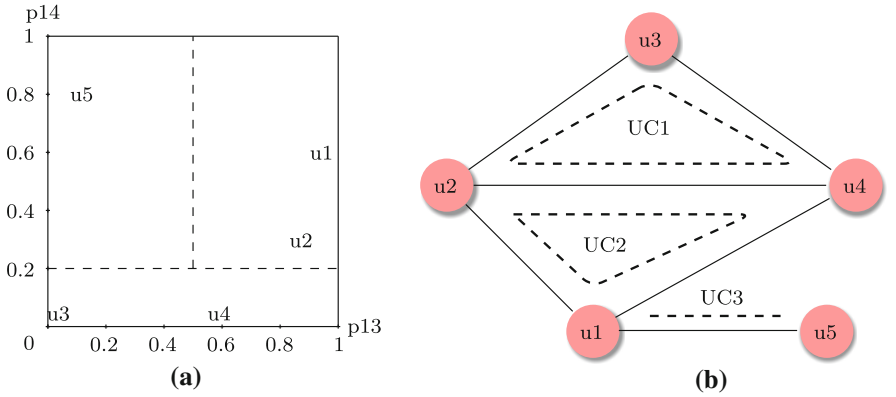

Fig. 3 Community discovery approaches. **a** Hard partitioning profiles and **b** Cliques in the user graph

have been used almost exclusively for the discovery of community models from the item graphs (e.g. Mobasher et al. 2002; Paliouras et al. 2000; Perkowitz and Etzioni 2000). Highly connected subgraphs of the item graphs naturally represent community models that can be used directly for personalization. In the clustering terminology used above, the objective that is maximized is the "similarity" among items in a community model $\mathbf{cm}_{ij} = T_{ij} \subset T_i$, i.e. $R_i(\mathbf{t}_{il}, \mathbf{t}_{im}), t_{il}, t_{im} \in T_{ij}$. The "dissimilarity" of items in different models is not usually maximized, leading often to highly overlapping sets. These methods are closely related to the popular *frequent itemset* data mining approach (Agrawal et al. 1993). According to this, user sessions are treated as customer transactions and the goal is to identify item subsets $T_{ij} \subset T_i$ (itemsets) that co-occur frequently in the transaction set, i.e. R_i is defined as a simple co-occurrence frequency. Frequent itemset approaches have been used widely for Web usage mining and community model discovery (Mobasher 2007). In contrast to clustering, these approaches emphasize the scalability to large sets of usage data and the flexibility of the discovery process. In some cases, specialized query languages have been developed for this purpose (Spiliopoulou and Faulstich 1998).

As exemplified in Fig. 2a, items define the dimensions of the space in which user clusters are sought. Similarly, item clusters are sought in the user space. Interestingly, the two views have also been combined to improve the community discovery process (e.g. George and Merugu 2005). Most frequently, the dimensions of the item space, i.e. all items, are too many and are thus reduced, in order to improve the computational performance of user community discovery. Beyond the discovery of item subsets (e.g. O'Connor and Herlocker 1999), traditional dimensionality reduction methods, such as Principal Component Analysis, have been employed for that purpose (e.g. Sarwar et al. 2000). According to these methods, the set of item dimensions T_i is mapped onto a reduced set of dimensions D_i that are linear combinations of T_i, i.e. each $d_{in} \in D_i$ is defined as $d_{in} = \sum_{m=1..|T_i|} w_{im} \times t_{im}, t_{im} \in T_i$. The aim of the dimensionality reduction algorithms is to achieve this compression with the minimum possible loss of information. As a result, the same communities can be discovered more efficiently in the reduced space defined by D_i.

Particularly interesting are the methods that associate T_i and UC_i probabilistically with D_i, i.e. each $t_{im} \in T_i$ and $uc_{il} \in UC_i$ has a probability of occurrence, $Pr(t_{im}|d_{in})$, $Pr(uc_{il}|d_{in})$, given a latent dimension $d_{in} \in D_i$. Due to the probabilistic association of users and items with latent dimensions, D_i can be thought as a set of communities. The probability $Pr(uc_{il}|d_{in})$ provides the membership of users to the community, while $Pr(t_{im}|d_{in})$ is a probabilistic model of the community. Several generative models have been proposed for these probabilities. Two of them, namely *probabilistic latent semantic analysis* (Hofmann 1999) and *latent Dirichlet allocation* (Blei et al. 2003) have been used for dimensionality reduction on usage data and subsequent discovery of communities (Hofmann 2004; Jin et al. 2004; Marlin 2003).

The use of probabilistic approaches to dimensionality reduction provides a natural modeling of overlapping communities. Each user and each item belongs in each latent community model to a certain degree, provided by the corresponding conditional probability. Probabilistic clustering methods adopt a similar model to probabilistic dimensionality reduction methods. Given a user community $UC_{ij} \subset UC_i$, each user $uc_{il} \in UC_i$ belongs to the community with a certain probability $Pr(uc_{il}|UC_{ij})$. Furthermore, each item $t_{im} \in T_i$ has a certain probability $Pr(t_{im}|UC_{ij})$ of belonging to the model of the community. These probabilities are estimated from data with the same methods that are used for probabilistic generative models mentioned above, e.g. the Expectation-Maximization algorithm or Gibbs sampling. In addition to probabilistic approaches (e.g. Ungar and Foster 1998), fuzzy clustering methods have also been used for the discovery of overlapping communities (e.g. Nasraoui et al. 2000). Fuzzy methods model user communities as fuzzy sets, in which each user belongs to a certain degree.

Given a probabilistic assignment of users to communities, a separate model can be constructed for each individual, as a function of all community models and the degree to which the user belongs in each community. This *community-based model* of the user, $\mathbf{ucm}_{il}$, is different from the user's profile, $\mathbf{uc}_{il}$, which stores the history of the user's interaction with the system. More specifically, for a user $uc_{il} \in UC_i$, and for a given clustering $\mathcal{UC}_i = \{UC_{ij}\}$, the model of the user is defined as a function of the models of the communities $\mathbf{CM}_i = \{\mathbf{cm}_{ij}\}$ and the corresponding probability assignments for the individual $\mathbf{Pr}_{il} = \{Pr(uc_{il}|UC_{ij})\}$, i.e. $\mathbf{ucm}_{il} = g_i(\mathbf{CM}_i, \mathbf{Pr}_{il})$. The function g_i can be defined in various ways:

- $\mathbf{ucm}_{il} = g_i(\mathbf{CM}_i, \mathbf{Pr}_{il}) = \mathbf{cm}_{ij}$, such that $argmax_j Pr(uc_{il}|UC_{ij})$, i.e. the model of the "closest" community to the user. Following the example of Sect. 3.2, this could be the set of pages {"p13","p14","p15"}, together with the corresponding probabilities estimated for the community.
- $\mathbf{ucm}_{il} = g_i(\mathbf{CM}_i, \mathbf{Pr}_{il}) = \mathbf{Pr}_{il} \times \mathbf{CM}_i$, i.e. a weighted mixture of all community models.
- $\mathbf{ucm}_{il} = g_i(\mathbf{CM}_i, \mathbf{Pr}_{il}) = \sum_{j=1}^{k} Pr(uc_{il}|UC_{ij}) \times \mathbf{cm}_{ij}$, such that $top\text{-}k_j$ $Pr(uc_{il}|UC_{ij})$, i.e. a weighted mixture of the k community models that are "closest" to the user.

A special case of the k closest models is to consider each user as a separate cluster, i.e. each community $UC_{ij} = \{uc_{ij}\}$ consists of a single user. In that case, the model of a user can be calculated as a mixture of the k closest communities, which are also in

this case individual users. This mixture can be weighted by the similarity of the users. In other words $\mathbf{ucm}_{il} = \sum_{j=1}^{k} R_i(\mathbf{uc}_{il}, \mathbf{uc}_{ij}) \times \mathbf{uc}_{ij}$, such that *top-k* $_j R_i(\mathbf{uc}_{il}, \mathbf{uc}_{ij})$. This is the *k-nearest neighbor* approach to community modeling, which has been particularly popular for collaborative filtering (e.g. Herlocker et al. 1999; Hill et al. 1995; Shardanand and Maes 1995). According to this approach, the aggregation of information from different users is only done locally and individually for each user. No pre-computation of clusters is necessary, although dimensionality reduction in the item space is often used for reasons of computational efficiency.

4 Community-based personalization

Collaborative filtering and recommendation have been the most widely-adopted types of personalization in commercial Web applications. According to this approach, a user is recommended items for viewing or purchase, based on the interest shown about these items by other similar users, i.e. the corresponding community. This is a general description of recommendation that can be specialized, according to the definition of the item and its use in the construction of the community model. Beyond collaborative recommendation, there are other types of Web personalization, where communities are of interest. The following categorization of personalization functions is proposed in Pierrakos et al. (2003):

Memorization: User salutation, Bookmarking, Personalized access rights.
Guidance: Recommendation of hyperlinks, User tutoring.
Customization: Personalized layout, Content customization, Customization of hyperlinks, Personalized pricing scheme, Personalized product differentiation.
Task performance support: Personalized errands, Personalized query completion, Personalized negotiations.

With the exception of the basic memorization functionality, user communities have been used in each of the other three main types of personalization. The rest of this section highlights various interesting aspects of applying user communities to the Web, in order to illustrate their potential for Web personalization.

Depending on the application, collaborative recommendation would be classified under the categories "guidance" or "customization" above. A large variety of products and services that are offered on the Web have been the objects of recommendation. The most well-known and widely-used example is movies, due to the early work of the GroupLens project[1] and the very useful datasets that the GroupLens team have made publicly available. The role of amazon.com and their early adoption of recommendation for their electronic market (Linden et al. 2003) has also been catalytic. Starting with books, they have made recommendation an integral part of their marketing strategy, which includes now a large variety of products. For each item being viewed by the site visitor, a range of other products are being advertised in a personalized manner. Thus, the electronic shop is being customized to the preferences of customer communities. The successful case of amazon.com has strongly affected the adoption of

[1] http://www.grouplens.org/.

recommendation by a variety of other businesses, ranging from music shops to travel agents and financial service providers. Thus, numerous successful applications have been presented in the literature (e.g. in Schafer et al. 2001; Wei et al. 2007). Due to its effectiveness, targeted personalized advertisement is now also widely used, even by search engines (Schroedl et al. 2010).

The nature of the items that get recommended plays an important role in the personalization process. In Sect. 2, we have assumed that items are only identified by a label and that no other information about them is available. In reality, though, this is hardly ever the case, as the items can have quite a rich description. For instance, in the classical example of movies, there are rich online descriptions of movies, e.g. in the Internet Movie Data Base.[2] The same is true for books and most of the other products of amazon.com and other online shops. Given the importance of this information, there are various proposals on how recommendation systems can make use of it:

Feature-level profiles: User profiles can be expressed at a sub-item level, e.g. the interest of a user about an actor, rather than a movie, can be stored in the user's profile. In many cases, where the item features are extracted from a database, this process is straightforward. However, there are items, for which feature extraction is non-trivial. For example, multimedia files for songs, images and movies require special software for the extraction of meaningful features.

Semantic item descriptions: The extraction of low-level features, particularly for multimedia data, is likely to make the user profiles incomprehensible and the data too sparse for community discovery algorithms. Therefore, a more condensed and meaningful representation of the items at a semantic level is preferable. This may take the form of meta-data annotations that are added manually to the multimedia items, e.g. the title of a movie or the genre of a song. Furthermore, recent research on the extraction of semantics from multimedia, on the basis of ontological descriptions, is very interesting (e.g. Naphade et al. 2006; Snoek et al. 2006). Ontologies provide a principled and machine-processable way to annotate content and are becoming increasingly popular in recommender systems (e.g. Burke 2007; Middleton et al. 2009; Mobasher et al. 2004). Especially since the advent of the semantic Web, several domain ontologies have been developed and are publicly available.[3] It is worth noting that ontologies formed the basis of user modeling, since the early stages of the field, when most of the proposed methods were driven by knowledge, rather than data (e.g. Rich 1979).

Semantic item relations: Beyond the semantic description of items, ontologies define relations among items. The simplest kind of such a relation is a taxonomy of item categories, which is common in electronic stores and other Web sites. This information can be used to enrich the user profiles, reduce the dimensionality of the data and even modify the manner in which the similarity between users and items is being estimated. Taxonomies can extend beyond individual Web sites to the whole

[2] http://www.imdb.com/.

[3] A search engine for ontologies is available at http://swoogle.umbc.edu/.

Web (Pierrakos and Paliouras 2010), while ontologies can provide non-taxonomic relations that are particularly interesting for personalization (Tao et al. 2007).

Beyond online businesses, user communities have been used for recommendation in a variety of other environments. Libraries (digital or not), museums and other cultural applications are among the most interesting ones (e.g. Ardissono et al. 2012; Kim and Fox 2004). In those environments, innovative interaction modes of the user with the system are often being researched, introducing various challenges to the recommendation process. Guiding tourists through the cultural points of interest in a city (Fink and Kobsa 2002) or museum visitors through the exhibits (van Hage et al. 2010) are just two such examples. In those cases, where the physical environment blends with the digital back-end, aspects such as the modeling of space and time become important for the recommendation process. Detailed tracking of the users' behavior can provide rich data for user modeling, such as where the users stop and for how long, what they choose to see and what to skip (Stock et al. 2007).

Furthermore, recommendation is not always driven by the interest of the user, but may be based on the user's expertise or special needs. Research on education systems is a good example where the role of interest is smaller, while the educational needs of the student and the targets of the teacher become much more important. User communities can play an important role in the guidance of the student through the teaching material and a variety of approaches have been proposed for the teaching of courses online (e.g. Anaya and Boticario 2009; Desmarais and Baker 2012; Mitrovic 2012; Zhuge 2009).

Educational systems are also interesting as an example of the overlap between the "guidance" and "task performance support" categories. This is because, much of the teaching process involves the completion of tasks and guiding the student often helps towards this goal. However, there is a number of other interesting task performance applications, where user communities can play an important role. Personalized search is one such example with great commercial interest, due to the continuous efforts of Web search companies to improve their search results. User communities can help in personalizing search results, often through the expansion of user queries (Almeida and Almeida 2004; Cui et al. 2003; Smyth 2007).

Finally, there are interesting applications of user community modeling that employ "customization", as a means to personalization. One of the early efforts in this direction was the development of *adaptive Web sites* (Perkowitz and Etzioni 2000). Most of this work has focussed on the selection of hyperlinks to propose to the user for further navigation through the Web site. This approach has been employed in a variety of personalized online systems, including educational ones (e.g. Farzan and Brusilovsky 2006; Zhuge 2009). It resembles collaborative recommendation, as the hyperlinks are recommended to the user. Another interesting attempt in the direction of "customization" was the personalization of product pricing, which however has raised a significant ethical controversy in the past.[4]

[4] http://www.wired.com/techbiz/media/news/2000/09/38622.

5 User communities in the social Web

In recent years, we have experienced the move from the "old Web" to the "social Web". The social Web has been facilitated by technological advances in the interaction of the users with Web resources (a.k.a. Web2.0 technologies) and has facilitated, in turn, two important social developments, termed *social media* and *social networks*. Social media or *user-generated content* represents the wide-spread participation of Web users in the generation of content, which has now become more volatile than ever before. On the other hand, *social networks* are typically Web sites that support the active networking of users, much in the spirit of Web-based communities, as presented in Sect. 2. Usually, social networks provide their users with the means to generate and share content and are thus considered the essence of the social Web. A well-known example of this combination of content generation and social networking is twitter.com, which is also among the most popular social Web sites. However, there is a wide variety of content that is nowadays shared using social Web applications, including pictures (flickr.com), video (youtube.com) and modern art (http://young.tate.org.uk/community).

Despite the fact that the ideas of social networks and social media are not new, the social Web has brought these ideas to the non-expert user. In fact, these extensions have widened considerably the user base of the Web to include people who had very little or no experience with computers. As a result, users who would stereotypically be content consumers have started contributing to the social Web, becoming *active Web users*.[5] By eliminating the distinction between consumers and producers, the social Web forces the redefinition of the different types of community, presented in Sect. 2, and introduces a new interesting type of community made of active users. Section 5.1 approaches *active Web user communities*, using the three types of community of Sect. 2 as a starting point, while Sect. 5.2 presents interesting challenges and opportunities for their discovery.

5.1 Active Web user communities

As briefly mentioned in Sect. 2, social networks can be considered the natural descendants of *Web-based communities* and community networks. As such, they bear similarities to those earlier approaches, like the goal of linking people with common interests or needs. A further similarity is that they are hosted by single sites, rather than linking different nodes of the Web. This is partly due to the business models of the social network sites and partly due to technological developments, such as cheap storage, that have facilitated the development of very large Web applications. However, social networks also have significant differences from their predecessors (Boyd and Ellison 2008), among which are:

[5] An alternative term that is often used is *prosumer*, combining the terms *producer* and *consumer*. We do not use this term here, because it is also used in a slightly different context to denote the *professional consumer*.

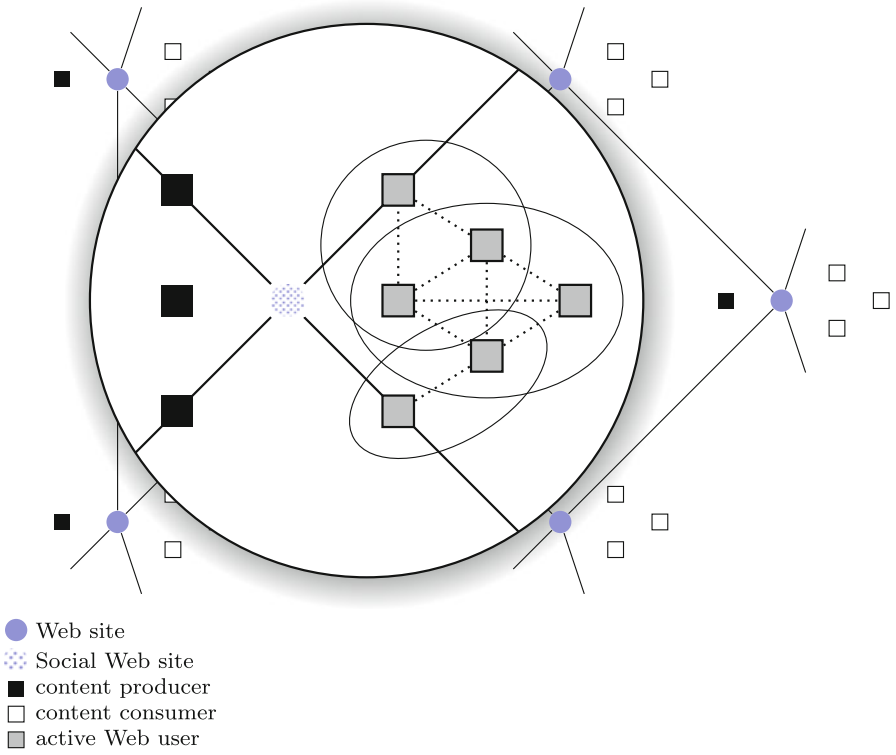

Fig. 4 Active Web user communities in a social Web site

- Their much larger user base.
- The diversity of their user base and their detachment from particular themes or geographic locations.
- The fact that people link to each other, but do not necessarily join predefined groups. A social network is a graph, in the sense of the Web itself, rather than a group of people.
- Their more participatory nature that turns passive consumers into active users, who provide content and information of many new and interesting types.

As a result, the role of active user communities is different from Web-based communities. If we assume that $s_i \in S$ is the Web node of a social network and UA_i is the set of active users of this network, a community of active users is a subset $UA_{ij} \subset UA_i$, rather than UA_i itself (Fig. 4). The set of active users UA_i is usually too large and diverse to be considered as a user community, on which to build personalization functionality.

Definition 5 (*Active Web User Community*) A subset $UA_{ij} \subset UA_i$ of active users of a Web site $s_i \in S$, usually a social network.

Due to the fact that social networks are graphs, a natural choice is to model user communities as *Web communities*, i.e. as subgraphs of the social network. Although the

analogy is technically very appealing, from the view of user modeling the two graphs are very different. Web communities, as per Definition 2, are subgraphs $S_j \subset S$ of the Web, where each individual node $s_i \in S_j$ is a Web site, created by a set of content producers UP_i and visited by a set of consumers UC_i. In social networks, the association of each node with a real user is much more direct and each community is a set of active users UA_{ij}. Being a true user community, this type of community is much more valuable to personalization than a Web community.

Thus, the definition of active user communities as sets of users of the social network leads naturally to the concept of Web user communities. However, there is an important new element in the social Web that makes active user communities much more interesting for personalization than traditional user communities. This is the manner in which users participate in the creation of the social network and the content therein. In the social Web, users provide much richer information about their preferences and needs, than what the logs of a traditional Web server could reveal. They choose their neighbors in the network, they publish their own content, they rate and tag content that other people have provided and participate in a number of online activities. Thus, the simple representation of user activity in terms of a user profile $\mathbf{uc}_{il}$ that stores the level of interest of the user for each of a set of items T_i is not sufficient any more. This change introduces a number of interesting challenges for user community discovery, as explained in Sect. 5.2 below.

At the same time, due to the rich interaction of the users with the system, the opportunities for personalization have also increased substantially. Recommendation of interesting items is but one of the many choices that can be explored. Personalization can affect the content that the user will consume or generate, the activities in which the user will participate and it can even affect the structure of the network, by suggesting potentially interesting new links. The discovery of active user communities can play an important role in this process.

5.2 Discovering active user communities

Section 5.1 has illustrated how the three very different types of community that were introduced in Sect. 2 come together in the context of the social Web to support the new concept of active Web user communities. It has also indicated that this development introduced challenges and opportunities for the research field of community discovery. In this subsection, we briefly review the most influential approaches to the discovery of active user communities and highlight some interesting aspects of the problem that have not been examined sufficiently so far.

Most of the work on community discovery in the social Web so far involves graph-based methods that have been used in the past for discovering either Web communities or Web user communities. They view the network as a graph of users, and try to identify efficiently the subgarphs, e.g. cliques or connected components, that suggest communities. Examples of such approaches can be found in several recent papers (e.g. Chen et al. 2010; Du et al. 2008; McDaid and Hurley 2010). The emphasis of this work is on the graph analysis method, which needs to be efficient, due to the large size of the graph, while at the same time producing useful communities. The same methods

are applicable to various different types of network, ranging from biological to telecommunication networks, where the notion of *community* plays an important role. An extensive survey of such methods can be found in Fortunato (2010).

The main issue with the graph-based approaches is that they tend to ignore all other information that is present in a social network, apart from the explicit links between users. One notable exception to this is the work on community detection in social tagging systems, also known as *folksonomies*, e.g. delicious.com. Community discovery in folksonomies is based on tri-partite associations between users, items and tags. This tri-partite view of the users already introduces challenges to the common community discovery methods that are based on simple two-dimensional user profiles (Table 2). The main technical challenge is the sparsity introduced in the data, as the dimensionality of the search space increases. The simple user-item matrices increase by one dimension, due to the tags added by users to the items. A number of methods have been proposed for community mining in folksonomies (Jäschke et al. 2007; Kashoob et al. 2010; Parra and Brusilovsky 2009; Siersdorfer and Sizov 2009) and a new interesting approach, called *tensor mining* has been developed (Sun et al. 2008; Symeonidis et al. 2008) for dealing with the tri-partite aspect of the problem. Tensors are multidimensional matrices that attempt to capture all aspects of the data, i.e. users, items and tags in the case of folksonomies. Data mining methods for tensors primarily attempt to reduce the dimensionality of the data, in order to make the datasets more dense and reveal significant patterns.

On the other side of the social Web, there are some methods that attempt to discover communities based on social media. Most of these methods have focused on Weblogs, treating them in many cases similarly to Web pages, i.e. building a graph of blogs that link to each other (Zhou and Davis 2007). However, there are also attempts to combine this basic network structure with other information that is provided by the active Web users, such as the exchange of comments and responses among them (Lin et al. 2007) or the interest shown by users for the content created by others (Seth et al. 2010). In a similar vein, the content generated in social networks has started being used to identify "missing links" in the network (e.g. Caragea et al. 2009).

These attempts illustrate the direction in which the growth of the social Web is pushing community discovery research. Traditional approaches are too limited, because they cannot model adequately the complex, multidimensional participation of active users in social networks and social media. The discovery of active user communities may be based on any and all of these dimensions of the users' activity. Thus, we need discovery methods that can work on structured data, which are also highly interconnected and potentially very large. The tensor methods, mentioned above, are one step in that direction. A different approach is the use of statistical relational learning methods (Xu et al. 2010) and probabilistic topic models (Pathak et al. 2008), which facilitate the discovery of communities on the basis of complex and potentially latent relations among users. Probabilistic topic models, such as probabilistic semantic analysis and latent Dirichlet allocation that were mentioned in Sect. 3.3, can be used to discover community models that bridge different aspects of the data, e.g. users and items, while relational learning methods support the refinement of existing knowledge about the domain, e.g. in the form of an ontology, with the analysis of relational data. Despite their appealing properties, though, these complex probabilistic models

are not suitable yet for the size of the data generated in social networks. Therefore, considerable research is still needed to turn them into practical tools for discovering communities of active Web users. One of the ways for restricting the complexity of the problem is the use of background knowledge, e.g. domain ontologies, which is naturally accommodated by relational models.

6 Concluding discussion

This article provided a quick overview of the use of communities in personalization in the past 20 years, positioning it in the broader context of Web community research. It has emphasized the importance of user community discovery methods in Web personalization and presented the main lines along which research in this field has progressed. Most importantly, it has revisited the definition of communities in the context of the social Web and proposed the concept of *active Web user communities*, as communities of users who participate actively in the creation of Web content and in related social activities. Given the socioeconomic importance of the move to the social Web, it is foreseen that this new type of community can shape much of the research in Web personalization in the coming years.

The discovery of active user communities has been researched in the past few years, albeit mostly with old tools that cannot address the complexity of the new environment. This article has presented some of the challenges that need to be tackled, in order to attain suitable community discovery methods. Most importantly, the multidimensional nature of user activity needs to be modeled adequately in the discovery of communities.

Beyond the need for new knowledge discovery methods, communities in the social Web, introduce new opportunities for personalization. Even the most common types of personalization, i.e. recommendation and filtering, can take several new forms in social networks and social media, such as the recommendation of new friends, the discovery of opportunities for content creation, the filtering of the prolific sources of user-generated content, etc. Furthermore, less popular types of personalization, such as tutoring and task performance support are likely to become more important, due to the diversity of complex social activities that take place in social networks, e.g. the various new types of online games. Thus, the social Web and its communities bring about new ideas for personalization services that can make the Web even more useful and interesting.

Among the expected developments that are likely to affect the formation and discovery of active user communities is also the move from single-site to multi-site social networks. An early step in this direction is the interconnection of popular networks and social media sites, e.g. using the OpenID[6] initiative. OpenID is a distributed and general authentication protocol that allows users to use the same id across different social Web services. As this interconnection becomes commonplace, the move from one site to another will become as easy as browsing the old Web, despite the requirement for registration. Therefore, active user communities across sites are likely to become

[6] http://openid.net/.

stronger and their discovery will provide new opportunities for personalization. One could claim that this is a step towards the *Web of People* (Berners-Lee 2000), where Web sites start disappearing and the Web starts taking the form of a human society. This is also strengthened by the disappearance of the computer itself and its replacement with a range of human-friendly mobile devices that blend ubiquitously with our everyday life. In this new environment, active user communities take the form of "traditional" communities, enabled and driven by communication and information technology.

Despite the excitement that these developments cause to both technologists and social scientists, skepticism is also growing quickly, raising a number of important issues. Privacy, trust and security are perhaps the most pressing of these issues. During the advent of social networks, we have witnessed an increased willingness to share private information that has surprised personalization researchers (Toch et al. 2012). This was partly due to the perceived reward for individuals who were willing to provide this information and partly due to the lack of awareness among many of the new users of the Web. There were also some technological advancements that have eased concerns, which however cannot catch up with the social change that the new Web is causing. This is evident by the number of illegal, offensive or simply annoying activities that are brought to light daily. Concerns are likely to increase further, as the integration of social Web services advances, e.g. different social networks sign agreements for data sharing. Privacy issues may thus become more serious and even damaging to the technological and social developments mentioned above. In most cases, this is an opportunity for new technologies and new business models, e.g. the recent Diaspora open-source project.[7] Being based on knowledge about the user, personalization is affected by these developments. Community-based personalization can be part of the solution to some of these problems, as it emphasizes the importance of social activity, rather than private information.

Another important source of concern is the special status of some users in the social networks. The role of power users and whether they act as facilitators or inhibitors to participation in the social Web is an issue of active research (Kittur et al. 2007; Nazir et al. 2008). Furthermore, the role of businesses, professionals and other organizations in some social networks is an issue (Kim 2010). These issues have been addressed in our society and as we move towards the Web of People, they are likely to find their on-line equivalents. Nevertheless, the effect of technology to this new society cannot be ignored as it may differentiate significantly any solution that will be sought. Again, the formation of communities, based on common characteristics, interests and goals, may help in overcoming some of these problems, much in the way in which it does in our society.

[7] http://www.joindiaspora.com.

References

Adamic, L.A., Adar, E.: Friends and neighbors on the web. Soc. Netw. **25**(3), 211–230 (2003)

Agrawal, R., Imielinski, T., Swami, A.N.: Mining association rules between sets of items in large databases. In: Buneman, P., Jajodia, S. (eds.) Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, pp. 207–216 (1993)

Almeida, R.B., Almeida, V.A.F.: A community-aware search engine. In: Feldman, S.I., Uretsky, M., Najork, M., Wills, C.E. (eds.) Proceedings of the Thirteenth International Conference on World Wide Web (WWW), New York, pp. 413–421 (2004)

Anaya, A.R., Boticario, J.: Clustering learners according to their collaboration. In: Borges, M.R.S., Shen, W., Pino, J.A., Barthès, J.P.A., Luo, J., Ochoa, S.F., Yong, J. (eds.) Proceedings of the 13th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Santiago, pp. 540–545 (2009)

Ardissono, L., Kuflik, T., Petrelli, D.: Personalization in cultural heritage: the road travelled and the one ahead. User Model. User-Adap. Inter. 22(1–2), 73–99 (2012)

Baeza-Yates, R.A., Ribeiro-Neto, B.A.: Modern Information Retrieval. ACM Press, Addison-Wesley (1999)

Berners-Lee, T.: Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web. Harper, San Francisco (2000)

Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)

Bouras, C., Igglesis, V., Kapoulas, V., Tsiatsos, T.: A web-based virtual community. Int. J. Web Based Comun. **1**(2), 127–139 (2005)

Boyd, D.M., Ellison, N.B.: Social network sites: definition, history, and scholarship. J. Comput. Mediat. Commun. **13**(1), 210–230 (2008)

Burke, R.D.: Hybrid web recommender systems. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) The Adaptive Web, Methods and Strategies of Web Personalization. Lecture Notes in Computer Science, vol. 4321, pp. 377–408. Springer, Heidelberg (2007)

Caragea, D., Bahirwani, V., Aljandal, W., Hsu, W.H.: Ontology-based link prediction in the livejournal social network. In: Bulitko, V., Beck, J.C. (eds.) Proceedings of the Eighth Symposium on Abstraction, Reformulation, and Approximation (SARA), Lake Arrowhead (2009)

Chen, W., Liu, Z., Sun, X., Wang, Y.: A game-theoretic framework to identify overlapping communities in social networks. Data Min. Knowl. Discov. **21**(2), 224–240 (2010)

Cooley, R., Mobasher, B., Srivastava, J.: Web mining: information and pattern discovery on the world wide web. In: Proceedings of the Ninth International Conference on Tools with Artificial Intelligence (ICTAI), Newport Beach, pp. 558–567 (1997)

Cooley, R., Mobasher, B., Srivastava, J.: Data preparation for mining world wide web browsing patterns. Knowl. Inform. Syst. **1**(1), 5–32 (1999)

Cui, H., Wen, J.R., Nie, J.Y., Ma, W.Y.: Query expansion by mining user logs. IEEE Tran. Knowl. Data Eng. **15**(4), 829–839 (2003)

Desmarais, M.C., Baker, R.S.J.d.: A review of recent advances in learner and skill modeling in intelligent learning environments. User Model. User-Adap. Inter. 22(1–2), 9–38 (2012)

Du, N., Wang, B., Wu, B.: Community detection in complex networks. J. Comput. Sci. Technol. **23**(4), 672–683 (2008)

Farzan, R., Brusilovsky, P.: Social navigation support in a course recommendation system. In: Wade, V.P., Ashman, H., Smyth, B. (eds.) Adaptive Hypermedia and Adaptive Web-Based Systems, Proceedings of the Fourth International Conference (AH). Lecture Notes in Computer Science, vol. 4018, pp. 91–100. Springer, Dublin (2006)

Fink, J., Kobsa, A.: User modeling for personalized city tours. Artif. Intell. Rev. **18**(1), 33–74 (2002)

Flake, G.W., Lawrence, S., Giles, C.L.: Efficient identification of web communities. In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Boston, pp. 150–160 (2000)

Fortunato, S.: Community detection in graphs. Phys. Rep. **486**, 75–174 (2010)

Gaudioso, E., Boticario, J.: User modeling on adaptive web-based learning communities. In: Palade, V., Howlett, R.J., Jain, L.C. (eds.) Proceedings of the Seventh International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES). Lecture Notes in Computer Science, vol. 2774, pp. 260–266. Springer, Oxford (2003)

George, T., Merugu, S.: A scalable collaborative filtering framework based on co-clustering. In: Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM), Houston, pp. 625–628 (2005)

Herlocker, J.L., Konstan, J.A., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), Philadelphia, pp. 230–237 (1999)

Hill, W.C., Stead, L., Rosenstein, M., Furnas, G.W.: Recommending and evaluating choices in a virtual community of use. In: Adelson, B., Dumais, S.T., Olson, J.S. (eds.) Proceedings of the Conference on Human Factors in Computing Systems (CHI), Denver, pp. 194–201 (1995)

Hofmann, T.: Probabilistic latent semantic analysis. In: Laskey, K.B., Prade, H. (eds.) Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI), Stockholm, pp. 289–296 (1999)

Hofmann, T.: Latent semantic models for collaborative filtering. ACM Trans. Inform. Syst. **22**(1), 89–115 (2004)

Jäschke, R., Marinho, L.B., Hotho, A., Schmidt-Thieme, L., Stumme, G.: Tag recommendations in folksonomies. In: Kok, J.N., Koronacki, J., de Mántaras, R.L., Matwin, S., Mladenic, D., Skowron, A. (eds.) Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD). Lecture Notes in Computer Science, vol. 4702, pp. 506–514. Springer, Warsaw (2007)

Jin, X., Zhou, Y., Mobasher, B.: Web usage mining based on probabilistic latent semantic analysis. In: Kim, W., Kohavi, R., Gehrke, J., DuMouchel, W. (eds.) Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Seattle, pp. 197–205 (2004)

Joachims, T., Freitag, D., Mitchell, T.M.: Web watcher: a tour guide for the world wide web. In: Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI), Nagoya, vol. 1, pp. 770–777 (1997)

Kashoob, S., Caverlee, J., Kamath, K.: Community-based ranking of the social web. In: Chignell, M.H., Toms, E. (eds.) Proceedings of the 21st ACM Conference on Hypertext and Hypermedia (HT), Toronto, pp. 141–150 (2010)

Kim, J.: User-generated content (ugc) revolution?: critique of the promise of youtube. Ph.D. thesis, University of Iowa (2010)

Kim, S., Fox, E.A.: Interest-based user grouping model for collaborative filtering in digital libraries. In: Chen, Z., Chen, H., Miao, Q., Fu, Y., Fox, E.A., Lim, E.P. (eds.) Digital Libraries: International Collaboration and Cross-Fertilization, Proceedings of the Seventh International Conference on Asian Digital Libraries (ICADL). *Lecture Notes in Computer Science*, vol. 3334, pp. 533–542. Springer, Shanghai (2004)

Kittur, A., Suh, B., Pendleton, B.A., Chi, E.H., Suh, B., Mytkowicz, T.: Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In: Presented at alt.CHI at ACM SIGCHI Conference on Human Factors in Computing Systems (CHI), San Jose, pp. 453–462 (2007)

Kohrs, A., Mérialdo, B.: Clustering for collaborative filtering applications. In: Proceedings of Computational Intelligence for Modelling, Control and Automation, Vienna (1999)

Konstan, J.A., Riedl, J.: Recommender systems: from algorithms to user experience. User Model. User-Adap. Inter. 22(1–2), 101–123 (2012)

Konstan, J.A., Miller, B.N., Maltz, D., Herlocker, J.L., Gordon, L.R., Riedl, J.: Grouplens: applying collaborative filtering to usenet news. Commun. ACM **40**(3), 77–87 (1997)

Kosala, R., Blockeel, H.: Web mining research: a survey. SIGKDD Explor. **2**(1), 1–15 (2000)

Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A.: Trawling the web for emerging cyber-communities. Comput. Netw. **31**(11-16), 1481–1493 (1999)

Lin, Y.R., Sundaram, H., Chi, Y., Tatemura, J., Tseng, B.L.: Blog community discovery and evolution based on mutual awareness expansion. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Silicon Valley, pp. 48–56 (2007)

Linden, G., Smith, B., York, J.: Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Comput. **7**(1), 76–80 (2003)

Marlin, B. Modeling user rating profiles for collaborative filtering. In: Thrun, S., Saul, L.K., Schölkopf, B. Advances in Neural Information Processing Systems (NIPS), Vancouver, British Columbia (2003)

McDaid, A., Hurley, N.: Detecting highly overlapping communities with model-based overlapping seed expansion. In: N. Memon, R. Alhajj (eds.) Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, pp. 112–119. Odense, Denmark (2010)

Middleton, S.E., Roure, D.D., Shadbolt, N.R: Ontology-based recommender systems. In: Staab, S., Studer, R. Handbook on Ontologies, International Handbooks on Information Systems, pp. 779–796. Springer, Heidelberg (2009)

Mitrovic, A.: Fifteen years of constraint-based tutors: what we have achieved and where we are going. User Model. User-Adap. Inter. 22(1–2), 39–72 (2012)
Mobasher, B.: Data mining for web personalization. In: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) The Adaptive Web, Methods and Strategies of Web Personalization. Lecture Notes in Computer Science, vol. 4321, pp. 90–135. Springer, Heidelberg (2007)
Mobasher, B., Cooley, R., Srivastava, J.: Automatic personalization based on web usage mining. Commun. ACM **43**(8), 142–151 (2000)
Mobasher, B., Dai, H., Luo, T., Nakagawa, M.: Discovery and evaluation of aggregate usage profiles for web personalization. Data Min. Knowl. Discov. **6**(1), 61–82 (2002)
Mobasher, B., Jin, X., Zhou, Y.: Semantically enhanced collaborative filtering on the web. In: Berendt, B., Hotho, A., Mladenic, D., van Someren, M., Spiliopoulou, M., Stumme, G. (eds.) Web Mining: From Web to Semantic Web, Revised Selected and Invited Papers of the First European Web Mining Forum, EMWF. Lecture Notes in Computer Science, vol. 3209, pp. 57–76. Springer, Heidelberg (2004)
Naphade, M.R., Smith, J.R., Tesic, J., Chang, S.F., Hsu, W.H., Kennedy, L.S., Hauptmann, A.G., Curtis, J.: Large-scale concept ontology for multimedia. IEEE MultiMed. **13**(3), 86–91 (2006)
Nasraoui, O., Frigui, H., Krishnapuram, R., Joshi, A.: Extracting web user profiles using relational competitive fuzzy clustering. Int. J. Artif. Intell. Tools **9**(4), 509–526 (2000)
Nazir, A., Raza, S., Chuah, C.N.: Unveiling facebook: a measurement study of social network based applications. In: Papagiannaki, K., Zhang, Z.L. (eds.) Proceedings of the Eighth ACM SIGCOMM Conference on Internet Measurement, Vouliagmeni, pp. 43–56 (2008)
O'Connor, M., Herlocker, J.L.: Clustering items for collaborative filtering. In: Proceedings of the ACM SIGIR Workshop on Recommender Systems: Algorithms and Evaluation, Berkeley (1999)
Orwant, J.: Heterogeneous learning in the doppelgänger user modeling system. User Model. User-Adap. Inter. **4**(2), 107–130 (1995)
Paliouras, G., Papatheodorou, C., Karkaletsis, V., Spyropoulos, C.D.: Clustering the users of large web sites into communities. In: Langley, P. (ed.) Proceedings of the Seventeenth International Conference on Machine Learning (ICML), Stanford, pp. 719–726 (2000)
Parra, D., Brusilovsky, P.: Collaborative filtering for social tagging systems: an experiment with citeulike. In: Bergman, L.D., Tuzhilin, A., Burke, R.D., Felfernig, A., Schmidt-Thieme, L. (eds.) Proceedings of the ACM Conference on Recommender Systems (RecSys), New York, pp. 237–240 (2009)
Pathak, N., DeLong, C., Banerjee, A., Erickson, K.: Social topic models for community extraction. In: Proceedings of the Second International Workshop on Advances in Social Network Mining and Analysis (SNAKDD), Las Vegas, pp. 77–96 (2008)
Pazzani, M.J., Muramatsu, J., Billsus, D.: Syskill & Webert: Identifying interesting web sites. In: Proceedings of the Thirteenth National Conference on Artificial Intelligence and Eighth Innovative Applications of Artificial Intelligence Conference (AAAI/IAAI), Portland, vol. 1, pp. 54–61 (1996)
Perkowitz, M., Etzioni, O.: Towards adaptive web sites: conceptual framework and case study. Artif. Intell. **118**(1-2), 245–275 (2000)
Pierrakos, D., Paliouras, G.: Personalizing web directories with the aid of web usage data. IEEE Trans. Knowl. Data Eng. **22**(9), 1331–1344 (2010)
Pierrakos, D., Paliouras, G., Papatheodorou, C., Spyropoulos, C.D.: Web usage mining as a tool for personalization: a survey. User Model. User-Adap. Inter. **13**(4), 311–372 (2003)
Rheingold, H.: The Virtual Community: Homesteading on the Electronic Frontier. Addison-Wesley, New York (1993)
Rich, E.: User modeling via stereotypes. Cogn. Sci. **3**(4), 329–354 (1979)
Sarwar, B.M., Karypis, G., Konstan, J.A., Riedl, J.: Application of dimensionality reduction in recommender system—a case study. Technical Report CS-TR 00-043, Computer Science and Engineering Department, University of Minnesota (2000)
Schafer, J.B., Konstan, J.A., Riedl, J.: E-commerce recommendation applications. Data Min. Knowl. Discov. **5**, 115–153 (2001)
Schroedl, S., Kesari, A., Neumeyer, L.: Personalized ad placement in web search. In: Proceedings of the 4th Annual International Workshop on Data Mining and Audience Intelligence for Online Advertising (AdKDD), Washington USA (2010)
Schuler, D.: Community networks: building a new participatory medium. Commun. ACM **37**(1), 38–51 (1994)
Seth, A., Zhang, J., Cohen, R.: Bayesian credibility modeling for personalized recommendation in participatory media. In: Bra, P.D., Kobsa, A., Chin, D.N. (eds.) Proceedings of the 18th International

Conference on User Modeling, Adaptation, and Personalization (UMAP). *Lecture Notes in Computer Science*, vol. 6075, pp. 279–290. Springer, Big Island (2010)
Shardanand, U., Maes, P.: Social information filtering: algorithms for automating "word of mouth". In: Katz, I.R., Mack, R.L., Marks, L., Rosson, M.B., Nielsen, J. (eds.) Proceedings of the Conference on Human Factors in Computing Systems (CHI), Denver, pp. 210–217 (1995)
Siersdorfer, S., Sizov, S.: Social recommender systems for web 2.0 folksonomies. In: Cattuto, C., Ruffo, G., Menczer, F. (eds.) Proceedings of the 20th ACM Conference on Hypertext and Hypermedia (HYPERTEXT), Torino, pp. 261–270 (2009)
Smyth, B.: A community-based approach to personalizing web search. IEEE Comput. **40**(8), 42–50 (2007)
Snoek, C., Worring, M., van Gemert, J., Geusebroek, J.M., Smeulders, A.W.M.: The challenge problem for automated detection of 101 semantic concepts in multimedia. In: Nahrstedt, K., Turk, M., Rui, Y., Klas, W., Mayer-Patel, K. (eds.) Proceedings of the 14th ACM International Conference on Multimedia, Santa Barbara, pp. 421–430 (2006)
Spiliopoulou, M., Faulstich, L.: WUM—a tool for www ulitization analysis. In: Atzeni, P., Mendelzon, A.O., Mecca, G. (eds.) Selected Papers of the International Workshop on World Wide Web and Databases (WebDB). Lecture Notes in Computer Science, vol. 1590, pp. 184–103. Springer, Heidelberg (1998)
Srivastava, J., Cooley, R., Deshpande, M., Tan, P.N.: Web usage mining: discovery and applications of usage patterns from web data. SIGKDD Explor. **1**(2), 12–23 (2000)
Staab, S., Angele, J., Decker, S., Erdmann, M., Hotho, A., Maedche, A., Schnurr, H.P., Studer, R., Sure, Y.: Semantic community web portals. Comput. Netw. **33**(1-6), 473–491 (2000)
Stock, O., Zancanaro, M., Busetta, P., Callaway, C., Kruger, A., Kruppa, M., Kuflik, T., Not, E., Rocchi, C.: Adaptive, intelligent presentation of information for the museum visitor in peach. User Model. User-Adap. Inter. **17**, 257–304 (2007)
Su, X., Khoshgoftaar, T.M.: A survey of collaborative filtering techniques. Adv. Artif. Intell. **2009**, 19p. doi:10.1155/2009/421425 (2009)
Sun, J., Tsourakakis, C.E., Hoke, E., Faloutsos, C., Eliassi-Rad, T.: Two heads better than one: pattern discovery in time-evolving multi-aspect data. Data Min. Knowl. Discov. **17**(1), 111–128 (2008)
Symeonidis, P., Nanopoulos, A., Manolopoulos, Y.: Tag recommendations based on tensor dimensionality reduction. In: Pu, P., Bridge, D.G., Mobasher, B., Ricci, F. (eds.) Proceedings of the ACM Conference on Recommender Systems (RecSys), Lausanne, pp. 43–50 (2008)
Tao, X., Li, Y., Zhong, N., Nayak, R.: Ontology mining for personalized web information gathering. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Silicon Valley, pp. 351–358 (2007)
Toch, E., Wang, Y., Cranor, L.F.: Personalization and privacy: a survey of privacy risks and remedies in personalization-based systems. User Model. User-Adap. Inter. 22(1–2), 203–220 (2012)
Ungar, L.H., Foster, D.P.: Clustering methods for collaborative filtering. In: Proceedings of the Workshop on Recommender Systems at the 15th National Conference on Artificial Intelligence, Madison (1998)
van Hage, W.R., Stash, N., Wang, Y., Aroyo, L.: Finding your way through the Rijksmuseum with an adaptive mobile museum guide. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) The Semantic Web: Research and Applications, Proceedings of the Seventh Extended Semantic Web Conference (ESWC), Part I. *Lecture Notes in Computer Science*, vol. 6088, pp. 46–59. Springer, Heraklion (2010)
Wei, K., Huang, J., Fu, S.: A survey of e-commerce recommender systems. In: Proceedings of the International Conference on Service Systems and Service Management, Tokyo, pp. 1–5 (2007)
Wu, K.L., Yu, P.S., Ballman, A.: Speedtracer: a web usage mining and analysis tool. IBM Syst. J. **37**(1), 89–105 (1998)
Xu, Z., Tresp, V., Rettinger, A., Kersting, K.: Social network mining with nonparametric relational models. In: Giles, C.L., Smith, M., Yen, J., Zhang, H. (eds.) Revised Selected Papers of the Second International Workshop on Advances in Social Network Mining and Analysis (SNAKDD). *Lecture Notes in Computer Science*, vol. 5498, pp. 77–96. Springer, Heidelberg (2010)
Yan, T.W., Jacobsen, M., Garcia-Molina, H., Dayal, U.: From user access patterns to dynamic hypertext linking. Comput. Netw. **28**(7–11), 1007–1014 (1996)
Zhou, Y., Davis, J.: Discovering web communities in the blogspace. In: Proceedings of the 40th Hawaii International Conference on Systems Science (HICSS), Waikoloa, p. 85 (2007)
Zhuge, H.: Communities and emerging semantics in semantic link network: discovery and learning. IEEE Trans. Knowl. Data Eng. **21**(6), 785–799 (2009)

Author Biography

Georgios Paliouras is a Senior Researcher and head of the Intelligent Information Systems division of the Institute of Informatics and Telecommunications at NCSR "Demokritos". He holds a Ph.D. and M.Sc. in Computer Science from the University of Manchester, UK, and a B.Sc. in Computing with Economics from the University of Central Lancashire, UK. His research has focused on knowledge discovery and machine learning methods and their application to user modeling, information extraction, text categorization, event recognition and ontologies. He has participated in several national and international research projects and has managed some of them. He has also served as board member in national and international scientific societies. Additionally, he is serving in the editorial board of international journals and has chaired international conferences. He is co-founder of the spin-off company i-sieve technologies Ltd.