Social networking meets recommender systems: survey

Guandong Xu

Advanced Analytics Institute,
University of Technology,
Sydney, Australia
Email: Guandong.Xu@uts.edu.au

Zhiang Wu*

Jiangsu Provincial Key Laboratory of E-Business,
Nanjing University of Finance and Economics,
Nanjing, China
Email: zawuster@gmail.com
*Corresponding author

Yanchun Zhang

School of Computer Science and Mathematics,
Victoria University,
Melbourne, Australia
Email: Yanchun.Zhang@vu.edu.au

Jie Cao

Jiangsu Provincial Key Laboratory of E-Business,
Nanjing University of Finance and Economics,
Nanjing, China
Email: Jie.Cao@njue.edu.cn

Abstract: Today, the emergence of web-based communities and hosted services such as social networking sites, wikis and folksonomies, brings in tremendous freedom of web autonomy and facilitate collaboration and knowledge sharing between users. Along with the interaction between users and computers, social media is rapidly becoming an important part of our digital experience, ranging from digital textual information to diverse multimedia forms. These aspects and characteristics constitute of the core of second generation of web. Social networking (SN) and recommender system (RS) are two hot and popular topics in the current Web 2.0 era, where the former emphasises the generation, dissemination and evolution of user relations, and the latter focuses on the use of collective preferences of users so as to provide the better experience and loyalty of users in various web applications. Leveraging user social connections is able to alleviate the common problems of sparsity and cold-start encountered in RS. This paper aims to summarise the research progresses and findings in these two areas and showcase the empowerment of integrating these two kinds of research strengths.

Keywords: social networking; recommender system; community detection; collaborative filtering; matrix factorisation; social recommender system.

Reference to this paper should be made as follows: Xu, G., Wu, Z., Zhang, Y. and Cao, J. (2015) 'Social networking meets recommender systems: survey', *Int. J. Social Network Mining*, Vol. 2, No. 1, pp.64–100.

Biographical notes: Guandong Xu is a Lecturer and Analytic Education Programme Leader in the Advanced Analytics Institute, University of Technology Sydney and he received his PhD in Computer Science from Victoria University, Australia. His research interests cover data mining, recommender systems and social network analysis. He has published three monographs in Springer and CRC press, and dozens of journal and conference papers including CJ, WWWJ, KAIS, KBS, IJCAI, WWW, AAAI, ICDM, and CIKM. He has been serving in editorial board or as guest editors for several international journals, such as the *Computer Journal*, *Journal of Systems and Software* and *World Wide Web Journal*, and he is the Assistant Editor-in-Chief of *World Wide Web Journal*.

Zhiang Wu is an Associate Professor of Jiangsu Provincial Key Laboratory of E-Business at Nanjing University of Finance and Economics, China. He received his PhD degree from Southeast University, China, in 2009. He is a member of the ACM, IEEE and CCF.

Yanchun Zhang is a Professor and the Director of Centre for Applied Informatics at Victoria University. He obtained his PhD in Computer Science from The University of Queensland in 1991. Since then, he has been an academic member at The University of Queensland, The University of Southern Queensland and Victoria University. He is an international expert in databases, data mining, health informatics, web information systems, and web services. He has published over 220 research papers in international journals and conferences proceedings, and authored/edited 12 books. His research has been supported by a number of Australian Research Council's project grants. His research has made some significant impacts on society. He is the Editor-in-Chief of *World Wide Web Journal*, and *Health Information Science and Systems Journal*. He is the Chairman of the International Web Information Systems Engineering Society (WISE Society).

Jie Cao is a Professor and the Director of Jiangsu Provincial Key Laboratory of EBusiness at Nanjing University of Finance and Economics. He received his PhD degree from Southeast University, China, in 2002. His main research interests include cloud computing, business intelligence and data mining. He has been selected in the Programme for New Century Excellent Talents in University (NCET).

1 Recommender systems

Nowadays, the internet has been well-known as a big data repository consisting of a variety of data types as well as a large amount of unseen informative knowledge, which can be discovered via a wide range of data mining or machine learning paradigms. Although the progress of the web-based data management research results in

developments of many useful web applications or services, like web search engines, users are still facing the problems of information overload and drowning due to the significant and rapid growth in amount of information and the number of users. In particular, web users usually suffer from the difficulties of finding desirable and accurate information on the web due to two problems of low precision and low recall caused by above reasons. For example, if a user wants to search the desired information by utilising a search engine such as Google, the search engine may provide the user not only the web content related to the query topic, but also a large mount of irrelevant information. It is sometimes hard for users to obtain their exactly needed information (Kosala and Blockeel, 2000; Chakrabarti, 2000) by using conventional search engines alone. Thus, the emerging of web has put forward a great deal of challenges to web researchers for web-based information management and retrieval. Web research academia is requested to develop more efficient and effective techniques to satisfy the increasing demands of web users, such as retrieving the desirable and related information (Hou and Zhang, 2003), creating good quality web community (Zhang et al., 2006; Kleinberg, 1998), extracting informative knowledge out of available information (Craven et al., 1998), capturing underlying usage pattern from web observation data (Srivastava et al., 2000), recommending or recommending user customised information to offer better internet service (Mobasher et al., 2002), and furthermore mining valuable business information from the common or individual customers navigational behaviour as well (Ghani and Fano, 2002).

Recommender systems (RSs), sometimes also called recommendation systems are a typical application to address information overload. RS could be viewed as a process that recommends the potentially interested items to users or predicts the possible ratings on various items for specific users based on their exhibited specific tastes or preferences. Since the items recommended or item ratings predicted are determined according to the personalised requirements, the term of RS is also co-occurred with another important terms from user modelling and human-computer interface, personalised or personalisation. To date, there are two basic kinds of approaches commonly used in RSs, namely content-based filtering and collaborative filtering systems (Dunja, 1996; Herlocker et al., 2004). Content-based filtering systems such as web watcher (Joachims et al., 1997) and client-side agent Letizia (Lieberman, 1995) usually generate recommendation based on the pre-constructed user profiles by measuring the similarity of web content to these profiles, while collaborative filtering systems make recommendation by referring other users' preference that is closely similar to current one. Recently, collaborative filtering-based approaches have been proved the most practical and successful methods in RS, evidenced by a number of commercial products or systems (Herlocker et al., 1999; Konstan et al., 1997; Shardanand and Maes, 1995).

Additionally, web usage mining has been recently proposed as an alternative method for not only revealing user access pattern, but also making web recommendation is the past decade (Mobasher et al., 2002). In the context of web usage mining, one important goal is to extract informative knowledge from web log files and identify underlying user functional interest that leads to common navigational activity. Basically, a user profile is created for representing a specific user navigational pattern based on mining usage data. Moreover, presenting the desired web content in a personalised style to user is carried out

by matching the current active user session with the discovered usage patterns. With the benefit of great progress in data mining research community, many data mining techniques, such as collaborative filtering based on the *k*-nearest neighbour (*k*NN) (Herlocker et al., 1999; Konstan et al., 1997; Shardanand and Maes, 1995), web user or page clustering (Mobasher et al., 2002; Han et al., 1998; Perkowitz and Etzioni, 1999), association rule mining (Agarwal et al., 1999; Agrawal and Srikant, 1994) and sequential pattern mining technique (Agrawal and Srikant, 1995) have been adopted in current web usage mining methods. Consequently, many efforts have been contributed and great achievements have been made in such research fields as web personalisation and recommendation systems (Lieberman, 1995; Joachims et al., 1997; Mobasher et al., 1999; Siaw et al., 1997), web system improvement (Cohen et al., 1998), web site modification or redesign (Perkowitz and Etzioni, 1998, 1999), and business intelligence and e-commerce (Büchner and Maurice, 1998).

As discussed above, the aim of RSs is to find the most matched user preference and behaviour patterns to the target user, via comparing the current user preference or behaviour with the existing observed user feedbacks or behaviours, or the recommendation models which are learned or trained from the available historic user behavioural data, and recommend a list of pages that user might be interested in. The former recommendation model is dependent on the explicit user behavioural data, whereas the latter process does heavily rely on the implicit model learned, where machine learning or data mining plays an important role. From the perspectives of optimisations, the explicit recommendation is an outcome of local optimal, instead, model-based approaches are more focused on the global optimised solutions. To perform recommendation efficiently and effectively, there are a variety of machine learning algorithms that have been well studied and developed, and can be used in web recommendation. In this section, we simply review several related algorithms that are often used in recommendation process.

1.1 kNN algorithm

*k*NN approach is the most often used recommendation scoring algorithm in many RSs, which is to compare the current user activity with the historic records of other users for finding the top *k* users who share the most similar behaviours to the current one. In conventional RSs, finding *k* nearest neighbours is usually accomplished by measuring the similarity in rating of items or visiting on web pages between current user and others. The found neighbouring users are then used to produce a prediction of items that are potentially rated or visited but not done yet by the current active user via collaborative filtering approaches. Therefore, the core component of the *k*NN algorithm is the similarity function that is used to measure the similarity or correlation between users in terms of attribute vectors, in which each user activity is characterised as a sequence of attributes associated with corresponding weights.

A variety of similarity functions can be used as measuring metrics. Among these measures, Pearson correlation coefficient and cosine similarity are two well-known and widely used similarity functions in RSs (Sarwar et al., 2001).

1.1.1 Correlation-based similarity

Pearson correlation coefficient, which is to calculate the deviations of users' ratings on various items from their mean ratings on the rated items, is a commonly used similarity function in traditional collaborative filtering approaches, where the attribute weight is expressed by a feature vector of numeric ratings on various items, e.g., the rating can be from 1 to 5 where 1 stands for the lest like voting and 5 for the most preferable one. The Pearson correlation coefficient can well deal with collaborative filtering since all ratings are on a discrete scale rather than on an analogous scale. The measure is described below. Given two users *i* and *j*, and their rating vectors and, the Pearson correlation coefficient is then defined by:

$$sim(i, j) = corr\left(R_i, R_j\right) = \frac{\sum_{k=1}^{n}\left(R_{i,k} - \overline{R}_i\right)\cdot\left(R_{j,k} - \overline{R}_j\right)}{\sqrt{\sum_{k=1}^{n}\left(R_{i,k} - \overline{R}_i\right)^2 \sum_{k=1}^{n}\left(R_{j,k} - \overline{R}_j\right)^2}} \tag{1}$$

where $R_{i,k}$ denotes the rating of user *i* on item *k*, $\overline{R}_i$ is the average rating of user *i*.

However, this measure is not appropriate in the web mining scenario where the data type encountered (i.e., user session) is actually a sequence of analogous page weights. To address this intrinsic property of usage data, the cosine coefficient is a better choice instead, which is to measure the cosine function of angle between two feature vectors. Cosine function is widely used in information retrieval research.

1.1.2 Cosine-based similarity

The cosine coefficient can be calculated by the ratio of the dot product of two vectors with respect to their vector norms. Given two vectors *A* and *B*, the cosine similarity is defined as:

$$sim(A, B) = \cos\left(\vec{A}, \vec{B}\right) = \frac{\vec{A}\cdot\vec{B}}{|\vec{A}|\times|\vec{B}|} \tag{2}$$

where '·' denotes the dot operation and '×' denotes the norm form.

1.2 Two main streams in RSs

Below we briefly discuss the two basic approaches in RSs, i.e., content-based RS and collaborative filtering-based RS.

1.2.1 Content-based recommendation

Content-based recommendation is a textual information filtering approach based on user's historic ratings on items. In a content-based recommendation, a user is associated with the attributes of the items that rated, and a user profile is learned from the attributes of the items to model the interest of the user. The recommendation score is computed by measuring the similarity of the attributes the user rated with those of not being rated, to determine which attributes might be potentially rated by the same user. As a result of

attribute similarity comparison, this method is actually a conventional information processing approach in the case of recommendation. The learned user profile reflects the long-time preference of a user within a period, and could be updated as more different rated attributes representing user's interest are observed. Content-based recommendation is helpful for predicting individual's preference since it is on a basis of referring the individual's historic rating data rather than taking other's preference into consideration.

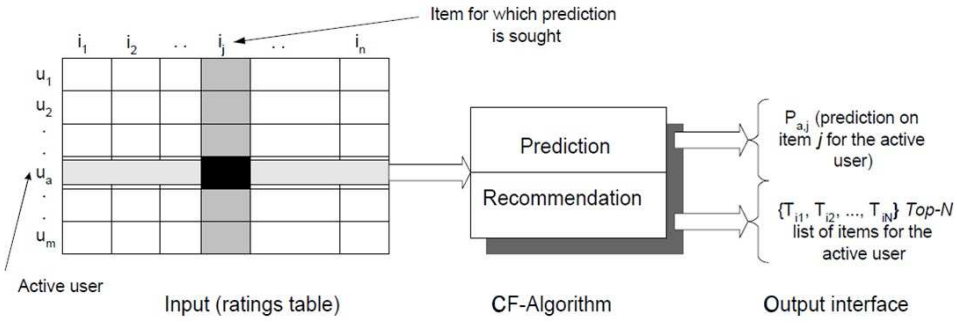
1.2.2 Collaborative filtering recommendation

Collaborative filtering recommendation is probably the most commonly and widely used technique that has been well developed for RSs. As the name indicated, collaborative RSs work in a collaborative referring way that is to aggregate ratings or preference on items, discover user profiles/patterns via learning from users' historic rating records, and generate new recommendation on a basis of inter-pattern comparison. A typical user profile in RS is expressed as a vector of the ratings on different items. The rating values could be either binary (like/dislike) or analogous-valued indicating the degree of preference, which is dependent on the application scenarios. In the context of collaborative filtering recommendation, there are two major kinds of collaborative filtering algorithms mentioned in literature, namely memory-based and model-based collaborative filtering algorithm (Sarwar et al., 2001; O'Conner and Herlocker, 1999; Herlocker et al., 2004).

Memory-based collaborative filtering

Memory-based algorithms use the total ratings of users in the training database while computing recommendation. These systems can also be classified into two sub-categories: user-based and item-based algorithms (Sarwar et al., 2001). For example, user-based *k*NN algorithm, which is based on calculating the similarity between two users, is to discover a set of users who have similar taste to the target user, i.e., neighbouring users, using *k*NN algorithm. After *k* nearest neighbouring users are found, this system uses a collaborative filtering algorithm to produce a prediction of top-*N* recommendations that the user may be interested in later. Figure 1 illustrates the framework and procedure involved in collaborative filtering.

Figure 1 The framework of collaborative filtering



Source: Sarwar et al. (2001)

Given a target user u, the prediction on item i is then calculated by:

$$p_{u,i} = \frac{\sum_{j=1}^{k} \left(R_{j,i} \cdot sim(u, j) \right)}{\sum_{j=1}^{k} sim(u, j)} \tag{3}$$

Here, $R_{j,i}$ denotes the rating on item i voted by user j, and only k most similar users (i.e., k nearest neighbours of user i) are considered in making recommendation. This kind approach is also called user-based kNN.

In contrast to user-based kNN, item-based kNN algorithm (Sarwar et al., 2001; Jin and Mobasher, 2003) is a different collaborative filtering algorithm, which is based on computing the similarity between two columns, i.e., two items. In item-based kNN system, mutual item-item similarity relation table is constructed first on a basis of comparing the item vectors, in which each item is modelled as a set of ratings by all users. To produce the prediction on an item i for user u, it computes the ratio of the sum of the ratings given by the user on the items that are similar to i with respect to the sum of involved item similarities as follows:

$$p_{u,i} = \frac{\sum_{j=1}^{k} \left(R_{u,j} \cdot sim(i, j) \right)}{\sum_{j=1}^{k} sim(i, j)} \tag{4}$$

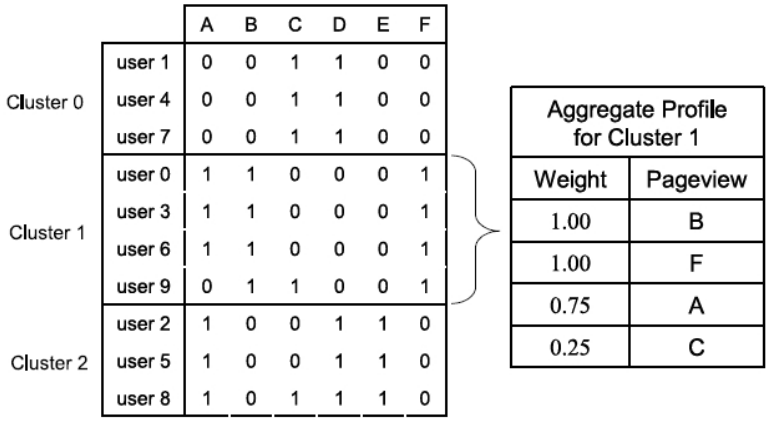
Here, $R_{j,i}$ denotes the prediction of rating given by user u on item j, and only the k most similar items (k nearest neighbours of item i) are used to generate the prediction.

Model-based recommendation

A model-based collaborative filtering algorithm is to derive a model from the historic rating data, and in turn, uses it for making recommendation. To derive the hidden model, a variety of statistical machine learning algorithms can be employed on the training database, such as Bayesian networks, neural networks, clustering and latent semantic analysis and so on. For example, in a model-based RS, named profile aggregations-based on clustering transaction (PACT) (Mobasher et al., 2002), clustering algorithm was employed to generate aggregations of user sessions, which are viewed as profiles via grouping users with similar access taste into various clusters. This is a process to assign data objects into various data groups or categories based on the similarity or distance between the objects such that the intra-group similarity within one group is maximised but the inter-group similarity is minimised. In the context of usage data, two types of clustering: clustering the transactions (or users) or clustering pageviews. After that the centroids of the user session clusters can be considered as access patterns/models learned from web usage data, in turn, used to make recommendation via referring to the web objects visited by other users who share the most similar access task to the current target user. Figure 2 depicts an example of using clustering to learn user profiles, in turn for web recommendation.

Although existence of different recommendation algorithms in RSs, it is easily found that these algorithms are both executing in a collaborative manner, and the recommendation score is dependent on the significance weight.

Figure 2 An example of deriving aggregate usage profiles from transaction clusters



Source: Mobasher et al. (2002)

Another example of model-based recommendation is implemented via *association rule mining*, which is an important data mining algorithm. Given a set of n pageviews, $P = \{p_1, p_2, \cdots, p_n\}$, and a set of m user transactions, $T = \{t_1, t_2, \cdots, t_m\}$, where each $t_i \in T$ is a subset of P.

Each transaction t as an l-length sequence of ordered pairs:

$$t = \left\{\left(p_1^t, w\left(p_1^t\right)\right), \left(p_2^t, w\left(p_2^t\right)\right), \cdots, \left(p_l^t, w\left(p_l^t\right)\right)\right\}$$

where each $p_i^t = p_j$ for some $j \in \{1, \cdots, n\}$, and $w(p_i^t)$ is the weight associated with pageview p_i^t in the transaction t.

Given a transaction T and a set $I = \{I_1, I_2, \cdots, I_k\}$ of frequent itemsets over T. The support of an itemset $I_i \in I$ is defined as

$$\sigma\left(I_i\right) = \frac{\left|\left\{t \in T : I_i \subseteq t\right\}\right|}{|T|} \tag{5}$$

An association rule r is an expression of the form

$$\mathrm{X} \Rightarrow \mathrm{Y}\left(\sigma_{\mathrm{r}}, \alpha_{\mathrm{r}}\right) \tag{6}$$

where X and Y are itemsets, $\sigma_{\mathrm{r}} = \sigma(\mathrm{X} \cup \mathrm{Y})$ is the support of $\mathrm{X} \cup \mathrm{Y}$ representing the probability that X and Y occur together in a transaction. The confidence for the rule r, σ_{r}, is given by $\sigma(\mathrm{X} \cup \mathrm{Y})/\sigma(X)$ and represents the conditional probability that Y occurs in a transaction given that X has occurred in that transaction.

For example, a high-confidence rule derived from the user purchase transaction data, such as {special-offers/, /products/software/} ⇒ {shopping-cart/} might provide some indication that a promotional campaign on software products is positively affecting online sales. Once such rules have been mined, apparently they could be used to make recommendations given the occurrence of predefined observations.

1.3 Advances in collaborative filtering

The two primary areas of collaborative filtering are the neighbourhood methods and latent factor models. Neighbourhood methods are focused on computing the similarities between items or users, e.g., the item-based approach determines a user's preference for an item based on ratings of 'neighbouring' items by the same user, whereas user-based approach calculates the preference degree on one specific by referring to like-minded users' rating. For example, given the movie *Saving Private Ryan*, its neighbours might include war movies, Spielberg movies, and Tom Hanks movies, or a mixture of those movies. As a result, the core of neighbourhood-based approaches is to accurately compute the similarity, which has been intensively discussed in the above section.

Konstas (2008) has systematically studies the various impacts on the rating predictions from the perspectives of bias, implicit feedback and item-item similarity.

1.3.1 Bias of users and items

Suppose the overall average rating is denoted by μ; the parameters b_u and b_i indicate the observed deviations of user u and item i, respectively, from the average. For example, suppose that you want an estimate for rating of the movie *Titanic*. The actual rating score is first determined by the average rating over all movies, μ. In addition, the score is also heavily dependent on the difference of individual user or movie (i.e., bias), e.g., the user is a fresh user or is a critical user, and so on. Thus, an estimate for an unknown rating is denoted by r_{ui} and accounts for the user and item effects: $r_{ui} = \mu + b_i + b_u$, where u means the global average, b_i means item bias, b_u means user bias.

1.3.2 Neighbourhood-based models

Besides the bias of users and items, the CF models should also involve other useful factors such as implicit feedback, item-item similarity, temporal dynamics and varying confidence level.

Implicit feedback

Implicit feedback can be used to gain insight into user preferences. Indeed, they can gather the behavioural information regardless of the user's willingness to provide explicit ratings. For example, an online shop can use its customers' purchases or browsing history to learn their prediction, in addition to the ratings those customers explicitly give. More clearly, we can use the implicit feedback (c_{ij}) to reflect the user preference. For two items i and j, an implicit preference by u to j leads us to modify our estimate of r_{ui} by c_{ij}, which is expected to be high if j is predictive on i.

Item-item similarity

The item-item similarity provides an additional hint to predict the rating. The hypothesis here is that the rating score will be boosted to the similar item. Given that the weight from j to i is denoted by w_{ij} (the similarity of item i and j) and will be learned from the data through optimisation, the following model describes each rating r_{ui} considering item-item similarity and implicit feedback by the equation:

$$\hat{r}_{ui} = u + b_i + b_u + \sum_{j \in R^k(i;u)} (r_{ui} - b_{ui}) w_{ij} + \sum_{j \in N^k(i;u)} c_{ij} \tag{7}$$

where $R^k(i; u) = R(u) \cap S^k(i)$ means the user's rating on i considering the most similar k items.

Neighbourhood models are most effective in detecting very localised closeness and they can rely on a few significant neighbourhood relations, but ignoring the vast majority of ratings by users. Consequently, these methods are unable to capture the totality of weak signals encompassed in all of a user's ratings.

1.3.3 Matrix factorisation models

As one of the most accurate single models for collaborative filtering, *matrix factorisation* (MF) (Konstas et al., 2009) is a Latent Factor model which is generally effective at estimating an overall structure hidden in the observations. In its basic form, MF characterises both items and users by vectors of factors inferred from item rating patterns. High relation between item and user factors indicates to a possible recommendation. For example, for movies, the discovered factors might measure obvious dimensions such as comedy versus drama, amount of action, or orientation to children. For users, each factor measures how much the user likes or dislike movies that score high or low on the corresponding movie categories. These methods have become popular in recent years by possessing a good scalability with satisfactory predictive accuracy. However, these approaches sometimes perform poorly at detecting strong adhesion among a small set of closely related items, precisely where neighbourhood models outperform.

The basic matrix of predicted ratings $R \in \mathbb{R}^{i_0 \times d}$, is modelled as:

$$\hat{R} = r_m + PQ^T \tag{8}$$

with matrices $P \in \mathbb{R}^{u_0 \times i_0}$ and $Q \in \mathbb{R}^{i_0 \times d}$, where u_0 denotes the number of users, and i_0 the number of items, d is the rank (or dimension of the latent space) with $d \leq i_0$, u_0, and $r_m \in \mathbb{R}$ is a global offset value. PQ^T reflects the interaction between users and items.

The major challenge is computing the mapping of each item and user to factor vectors P, Q. After the RS completes this decomposition, it can easily estimate the rating a user will give to any item by using equation (8).

Temporal dynamics

In reality, product branding and popularity constantly change as new merchandises launched. Similarly, customers' preference or taste evolve with the time, impact them to making decisions. Thus, the MF should take the temporal effects, i.e., the dynamic, time-drifting nature of user-item interactions into account.

$$\hat{R}_{ui}(t) = \mu + b_i(t) + b_u(t) + q_i^T p_u(t) \tag{9}$$

The first temporal effect addresses the fact that an item's popularity might change over time, while the second temporal effect allows users to change their baseline ratings over time. For example, a user who tended to rate an average movie '4 stars' might now rate such a movie '3 stars'.

1.3.4 Combined models

Neighbourhood models are most effective at detecting very localised closeness and they rely on a few significant neighbourhood relations to make recommendations. Given that the its close neighbours are able to provide reliable and accurate ratings, this approach is very effective and efficient. However, these methods fail to capture the totality of weak signals encompassed in all of a user's ratings. Latent factor models are generally effective at estimating an overall preference based on revealing the hidden topical relations of users and performance. However, these models might not be applicable in the cases that there are some strong associations among a small set of closely related items, where neighbourhood models are mostly appropriate. Intuitively, combining MF with neighbourhood into a unified framework will undoubtedly increase the prediction accuracy by benefiting the advantages of both neighbourhood and latent factor approaches.

$$\hat{r}_{ui} = u + b_i + b_u + \frac{\sum_{j \in R^k(i;u)} (r_{ui} - b_{ui}) w_{ij}}{\sqrt{|R^k(i;u)|}} + \frac{\sum_{j \in N^k(i;u)} c_{ij}}{\sqrt{|N^k(i;u)|}} + q_i^T p_u \tag{10}$$

In a sense, equation (10) provides a three-component model for recommendations. The first component, $\mu + b_u + b_i$ describes the general properties of the item and the user, without accounting for any involved interactions. The next component takes the neighbourhood and item-item similarity into account. The final $q_i^T p_u$ provides a latent correspondence between users and items. This combined framework lays down a reference model to accommodate various factors that may influence the MF, which is now widely used in other social RSs. To estimate P and Q, some statistical learning algorithms are adopted (Konstas, 2008).

2 Social networking

The all-round infiltration of Web 2.0 these days further promotes the development of social network which have already brought massive changes to society. The most distinctive characteristic of the social network is that a user of social media can be both a consumer and a producer (Tang et al., 2010). As hundreds of millions of users abandon themselves to various social media, i.e., blog, microblog, Wiki, everybody can be a media outlet. Absorbing a large amount of users is an another distinctive characteristic of the social network. For instance, more than 800 million users have registered in Facebook being the third largest society in the world just behind China and India.

A social network can often be modelled as a graph, of which the nodes are called actors (individuals or organisations) and the edges connecting nodes represent various ties. It is interesting to note that the tie might be in one or more types of interdependencies including shared values, visions and ideas; social contacts; kinship; conflict; financial exchanges; trade; joint membership in organisations; and group participation in events, among numerous other aspects of human relationships (Serrat, 2009). Social network analysis (SNA) assumes relationships are important, and thus aims to map and measure relationships to uncover what facilitates or impedes the knowledge flows that bind interacting units, e.g., who knows whom, and who shares what information with whom by what type of media. SNA continues to intrigue numerous researchers and theorists from many fields including sociology, physics, intelligent analysis, epidemiology, targeted marketing and RSs, and so on.

Below, we will introduce several research branches and their state-of-the-art in social networking (SN).

2.1 Community detection in social networks

Community detection divides actors in a network into groups that are meaningful, useful, or both. So far, there is no standard definition of the network community. Generally, a network community refers to a group of actors within which the connecting links are dense but between which they are sparse. These closely-knit groups that the actors in a network tend to be are also called *communities*, *clusters*, *cohesive subgroups* or *modules* in different contexts (Tang et al., 2010).

Community detection has become a fundamental problem ever since the network science came into vogue. In the literature, the existing community detection methods can fall into two categories, one is with global models and the other is not. Another burgeoning subfield in community detection is *community extraction*. In what follows, we will briefly review two kinds of community detection methods and the recent work on community extraction.

2.1.1 Community detection with global models

The methods with global models typically consider the global topology of a network, and aim to optimise a criterion defined over a network partition. Some methods along this line include the Kernighan and Lin algorithm (1970), latent space models (Handcock et al., 2007), stochastic block models (Karrer and Newman, 2011), modularity optimisation (Newman, 2004), and traditional clustering techniques (Slater, 2008) such as K-means, multi-dimensional scaling (MDS), and spectral clustering. The differences between these methods ultimately come down to the precise definition of a ‘denser’ community, i.e., the global criterion and the algorithmic heuristic followed to identify such sets.

2.1.2 Community detection without global models

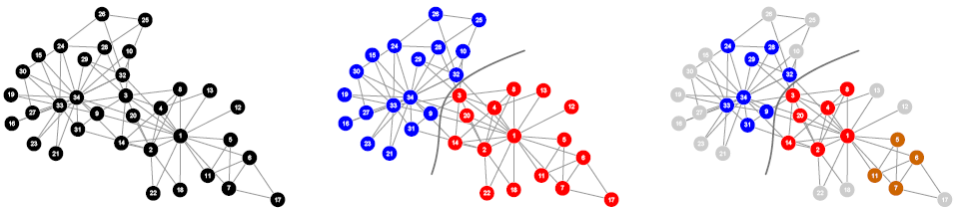
The methods without global models typically employ a bottom-up strategy to find communities. They often start by defining the properties of a node, a pair of nodes, or a group of nodes in a same community, and then search within a whole network for the communities that hold the proposed properties (Tang et al., 2012). A network’s global

community structure is detected by considering the ensemble of communities obtained by looping over all of these local structures. For example, the method of *k*-clique percolation (Palla et al., 2005) is based on the concept of *k*-clique, and a *k*-clique community is then defined as the union of all 'adjacent' *k*-cliques, which by definition share *k*–1 nodes. Besides *k*-clique, a community could be regarded as a clique, a *k*-club (Mokken, 1979), a quasi-clique (Abello et al., 2002), an equivalent structure, or the combination of node pairs that have nodes similar to each other, as measured by for example, Jaccard coefficient or cosine similarity (Tang et al., 2010).

2.1.3 Community extraction

The problem of community detection is, a real-life network might probably contain nodes that have weak connections to any communities. So grouping these weakly connected nodes with tighter communities actually impedes us from finding the genuine communities. Therefore, recently researchers present community extraction to extract genuine communities *hidden* inside the massive networks. Figure 3 depicts an example on the `Karate-Club` network to illustrate the difference between community detection and community extraction. Community detection splits up all nodes into several groups as shown in Figure 3(b), while community extraction removes the periphery nodes [the grey nodes in Figure 3(c)] and partitions the core nodes into several groups. Borgatti and Everett (1999) tried to divide nodes into core and periphery sets based on the proposed CP measure, but their methods can only work for small-size networks. Local community detection (Clauset, 2005) aims to find the tightest community around a given node locally rather than globally. Recently, Zhao et al. (2011) proposed a criterion *W* to extract tight communities one by one, but the tabu search prevents it from being further used for large-scale networks. Some hierarchical models also seek to highlight communities by excluding unrelated nodes (Clauset et al., 2008).

Figure 3 Illustration of the difference between community detection and community extraction on the `Karate-Club` network, (a) original network (b) community detection (c) community extraction (see online version for colours)



(a) (b) (c)

2.1.4 Community evaluation

Evaluation measures and methods are important to the comparison of different community detection/extraction methods. For some small-scale networks with ground truth, the community label of every actor is known. This ideal case often occurs on the synthetic data, or some well-studied tiny networks, i.e., `Karate-Club`, `American`

`college football`, `political blogs`, and `political books`, etc.[1] A multitude of validation measures from clustering field (Wu et al., 2009), such as normalised mutual information (NMI), normalised Rand index (R_n), variation of information (VI), etc., can be utilised to compare the ground truth with identified communities.

However, there hardly exist community labels for real-world large-scale networks like `Epinions`, `Slashdot`, `Enron`, `citation networks`, etc.[2] We usually have two common ways to evaluating such networks. The first one employs modularity Q as an internal measure. It is computed as follows:

$$Q = \sum_{i=1}^{K} \left(\frac{|E|_{c_i}}{|E|} - \left(\frac{\sum_{x \in c_i} \deg[x]}{2|E|} \right)^2 \right), \tag{11}$$

where K is the number of clusters, c_i denotes cluster i, $|E|_{c_i}$ is the edges in c_i and deg[x] is the degree of node x. The value of Q is in the interval: (–1, 1), and a larger value indicates a better partitioning result. Another one is to capture the semantics of identified communities. For example, tags or topics of users in the same community can reflect their common interest, which can in turn validate the communities detected from network topology.

2.1.5 Summary

Since the nature of network communities is not very clear now, more and more attentions have been paid to deal with this interesting and challenging problem. To elaborate all the related work on community detection is definitely impossible. Readers with this interest may refer to some excellent books and survey papers (Fortunato, 2010; Tang et al., 2010).

2.2 Temporal analysis on social networks

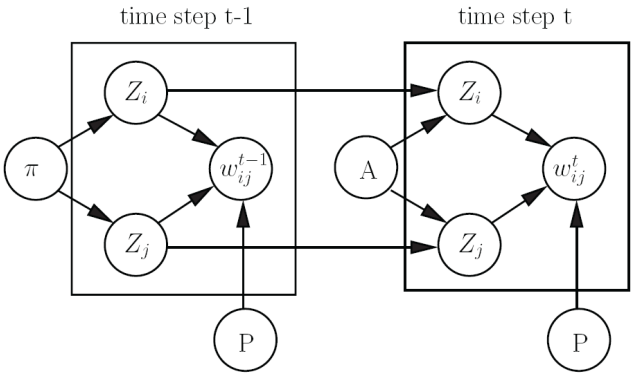
Most of the work on the SNA treats the network as a *static network*, where the static network is either derived from aggregation of data over all time or intercepted from a snapshot of time series data (Lin et al., 2009). It is obvious that many ingredients of social network, such as links, communities, public opinions, and sentiments of users, etc., are varying with the passing of time. Kossinets and Watts (2006) empirically observed some real-world networks were evolving. Time stamps on social networks add a temporal dimension to the data, and thus analysis of the social networks with this temporal data can lead to new insights into the system.

The *temporal social network* is usually defined a sequence of snapshot graphs indexed by time (Lin et al., 2009; Pietilainen and Diot, 2012). Every snapshot is a static network, $G_t(V, E_t)$, where V denotes a set of actors and E_t denotes the relationship between two actors during a time window $[t, t + 1)$. In this section, we will review several subfields based on the dynamic view of social networks.

2.2.1 Community evolution

Community evolution processes the temporal social network to produce a sequence of communities; that is a community for each timestep. Most of the existing studies deal with the dynamic *single-mode* network in which only one type of actors are present. Many models with heuristics have been proposed for this problem, including dynamic stochastic block model (DSBM) (Yang et al., 2009), approximation algorithms by casting it as a graph colouring problem (Tantipathananandh et al., 2007), information-theoretic-based method (Sun et al., 2007), and (hierarchical) Dirichlet process-based method (Xu et al., 2008a, 2008b), etc.

Figure 4 Graph representation of DSBM



Source: Yang et al. (2009)

We use the DSBM (Yang et al., 2009) as an example to illustrate the internal factors and their relationships in community evolution. Let $W^{(t)}$ be the snapshot of a social network at timestep t, each element ω_{ij} be a binary number indicating the presence and absence of an edge between nodes i and j, and $\mathcal{W}_T = \{W^{(1)}, W^{(2)}, \cdots, W^{(T)}\}$ be a sequence of snapshots over T discrete timesteps. Corresponding to $\mathcal{W}_T$, we define $\mathcal{Z}_T = \{Z^{(1)}, Z^{(2)}, \cdots, Z^{(T)}\}$ to denote the sequence of community assignments, where $Z^{(t)}$ is a matrix of which each element $z_{ik}^{(t)}$ denotes the node i is assigned to k^{th} community at timestep t. DSBM then estimates the parameters $\theta = \{\pi, P, A\}$ from the historical data, where $\pi_i = \{\pi_1, \cdots, \pi_K\}$ is the initial probability for i^{th} node to be assigned to K communities, P is a matrix of which each element P_{kl} is the parameter of a Bernoulli distribution followed by a generated link from k^{th} to l^{th} community, and A is a transition matrix of which each element A_{kl} is the probability a node change from k^{th} to l^{th} community. Figure 4 shows the probabilistic graphical model of DSBM. Two learning models, i.e., *offline learning and online learning*, are presented. The offline learning tries to learn the community assignments of all nodes at all timesteps [see equation (12)], while the online learning only tries to learn the community assignment at timestep t [see equation (13)].

$$\mathcal{Z}_T^* = \arg\max_{\mathcal{Z}_T} P\left(\mathcal{Z}_T \mid \mathcal{W}_T\right) = \arg\max_{\mathcal{Z}_T} P\left(\mathcal{W}_T, \mathcal{Z}_T\right), \tag{12}$$

$$Z^{*(t)} = \arg\max_{Z} P\left(Z^{(t)} \mid W^{(t)}, Z^{(t-1)}\right), \tag{13}$$

where $P(\mathcal{W}_T, \mathcal{Z}_T) = \int P(\mathcal{W}_T, \mathcal{Z}_T \mid \theta)\Pr(\theta)d\theta$. In Yang et al. (2009), a Gibbs sampling-based method is employed to optimise the posterior probabilities in both offline and online learning algorithms. Besides the aforementioned works on single-mode networks, some recent works (Tang et al., 2012b, 2012c) turn to solve community evolution in *multi-mode network*, i.e., heterogeneous network, involving more than one type of actors and interactions.

2.2.2 Link prediction

The *link prediction* problem can be formalised as Liben-Nowell and Kleinberg (2007): given a snapshot of a social network at time t, we seek to accurately predict the edges that will be added to the network during the interval from time t to a given future time t'. The existing link predictors can be divided into two types according to the information used for prediction (Kashima et al., 2009):

1 topological-information-based link predictors

2 node-information-based link predictors.

The former type was constructed by using graph theoretic measures such as the number of common neighbours, Katz measure (Liben-Nowell and Kleinberg, 2007), and mixed graph-theoretic measures (Leroy et al., 2010). The latter type takes node information as features, and thus treats the link prediction as a binary classification problem. Most of the works adopt the supervised learning method to train the classifier (Hasan et al., 2006; Wang et al., 2007; Doppa et al., 2009), while some recent works employ the semi-supervised learning classification (Kashima et al., 2009; Raymond and Kashima, 2010).

2.2.3 Topic evolution

The aforementioned community evolution and link prediction actually investigate the evolution of network topology. Beside the topological structure, SN also provides us abundant *social documents*; that is, users can freely release their opinions and sentiments in terms of blogs, tweets, microblogging, etc. Mining temporal social documents to interpret and understand human behaviours is becoming an important direction of SN. Many studies focus on detecting emerging topics and observing their trends (Boykin and Merlino, 2000; Kleinberg, 2002), while other works try to summarise the complete evolutionary topics (Mei and Zhai, 2005; Chen and Chen, 2012). Recently, the increasing attention has been paid to mining stories, i.e., the discussion topics lasting for a limited time duration, from blogs and tweets (Qamra et al., 2006; Meng et al., 2012).

2.2.4 Spatio-temporal analysis

In addition to the temporal dimension, many works have also taken spatial dimension into the analysis of social networks, and thus investigated how these additional dimensions (i.e., space and time) influence the structural properties and the dynamics behaviour of networks. How spatio-temporal data can be used to infer social ties (David et al., 2010; Eagle et al., 2010; Cranshaw et al., 2010), and how social ties influence the mobility of users (Backstrom et al., 2010; Cho et al., 2011) are the new hot spots. More broadly,

conducting spatio-temporal analysis on social networks has many applications, such as location-based recommendations (Ge et al., 2010; Zheng et al., 2011), urban planning (Glaeser and Kahn, 2004), and spread of diseases (Eubank et al., 2004; Culotta, 2010).

2.2.5 Summary

Temporal analysis on social networks entrenches upon the management and mining on the spatio-temporal data, and crosses with many disciplines to facilitate a sea of fascinating applications. The cited chapter and book (Hasan and Zaki, 2011; Zheng and Zhou, 2011) are good guides on this direction.

2.3 Social influence analysis

Social influence refers to the behavioural change of individuals affected by others in a network. To quantitatively measure the strength of social influence, lots of influence related statistics have been presented. Among them, edge measures capture the tie strength on a pair of actors. Commonly, if the overlap of neighbourhoods between two actors is large, they are considered to have a strong tie (Granovetter, 1973). The local/global bridge is presented to describe the weak-tie nodes (Granovetter, 1973), that is, when there is no overlap, the connection of two nodes is a bridge. Node measures, such as degree, centrality, and betweenness, are defined to measure the importance of a node in the network. In what follows, three interesting aspects of social influence analysis will be reviewed, including social similarity and influence (i.e., homophily), information propagation and maximisation.

2.3.1 Homophily

It is the basic feature to consider the central problem for social influence, i.e., the relationship between influence and correlation. Homophily suggests that an actor in the social network tends to be similar to their connected neighbours or 'friends'. Singla and Richardson (2008) conducted a large-scale experiments to confirm the existence of homophily. The homophily results from two main factors *selection* and *social influence*. For example, RSs can be explained as selection based on similarity, and information propagation and influence maximisation can be explained as social influence. Many studies are then devoted to quantify selection and influence. Scripps et al. (2009) proposed the formal computational definitions of selection (S) and influence (I) as follows:

$$S = \frac{p\left(a_{ij}^{t} = 1 \mid p\left(a_{ij}^{t-1} = 0, \left\langle \mathbf{x}_i^{t-1}, \mathbf{x}_j^{t-1} \right\rangle > \epsilon\right)\right)}{p\left(a_{ij}^{t} = 1 \mid p\left(a_{ij}^{t-1} = 0\right)\right)} \tag{14}$$

$$I = \frac{p\left(\left\langle \mathbf{x}_i^{t}, \mathbf{x}_j^{tT} \right\rangle > \left\langle \mathbf{x}_i^{t-1}, \mathbf{x}_j^{t-1} \right\rangle \middle| a_{ij}^{t-1} = 0, a_{ij}^{t} = 1\right)}{p\left(\left\langle \mathbf{x}_i^{t}, \mathbf{x}_j^{tT} \right\rangle > \left\langle \mathbf{x}_i^{t-1}, \mathbf{x}_j^{t-1} \right\rangle \middle| a_{ij}^{t-1} = 0\right)}, \tag{15}$$

where the denominator of equation (14) is the conditional probability that an unlinked pair will become linked, and the numerator of equation (14) is the same probability for unlinked pairs whose similarity exceeds the threshold ϵ, the denominator of equation

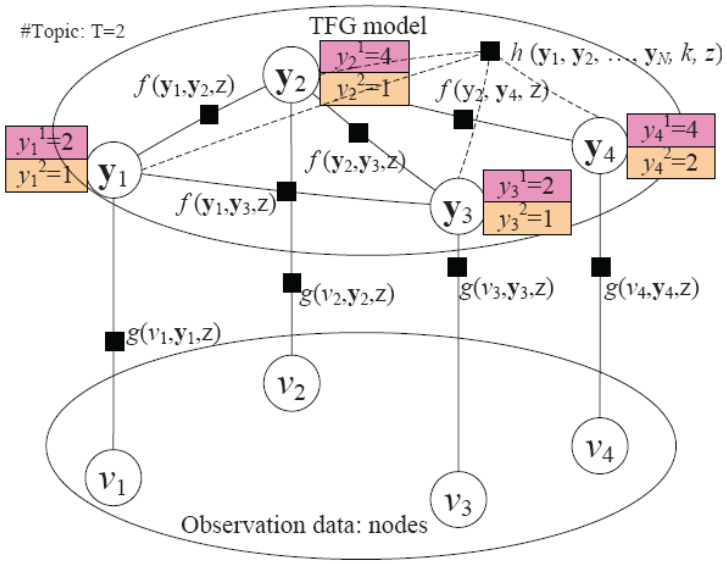
(15) is the conditional probability that similarity increases from time $t-1$ to t between two nodes that become linked at time t, and the numerator of equation (15) is the same probability for two nodes who were not linked at time $t-1$. Based on this definition, a matrix alignment framework was presented to learn the weight of different attributes for establishing relationships between users, which can be done by optimising the following objective function.

$$\min_{W} \sum_{t=1}^{T} \left\| A^{t} - X^{t-1} W X^{(t-1)\top} \right\|_{F}^{2}, \tag{16}$$

where the diagonal elements of W denote the weights of attributes and $\|\cdot\|_F$ denotes the Frobenius norm. The above work does not differentiate the influence from different topics. To remedy this, Tang et al. (2009) propose a topical factor graph (TFG) model as shown in Figure 5 to formalise the topic-level social influence analysis. In TFG model, given a network with N nodes, $\{v_i\}_i^N$ is a set of observed variables and $\{\mathbf{y}_i\}$ is a set of hidden vectors. Then, three types of feature functions are defined to capture the network information: node feature function $g(v_i, \mathbf{y}_i, z)$, edge feature function $f(\mathbf{y}_i, \mathbf{y}_j, z)$, and global feature function $h(\mathbf{y}_1, \cdots, \mathbf{y}_N, k, z)$. The task of social influence is cast as that of identifying the node which has the highest probability to influence another node on a specific topic along with the edge. This is the same as that of maximising the following objective function:

$$P(\mathbf{v}, \mathbf{Y}) = \frac{1}{Z} \prod_{k=1}^{N} \prod_{z=1}^{T} h(\mathbf{y}_1, \cdots, \mathbf{y}_N, k, z) \prod_{i=1}^{N} \prod_{z=1}^{T} g(v_i, \mathbf{y}_i, z) \prod_{e_{kl} \in E} \prod_{z=1}^{T} f(\mathbf{y}_k, \mathbf{y}_l, z) \tag{17}$$

Figure 5 Graph representation of the TFG model (see online version for colours)



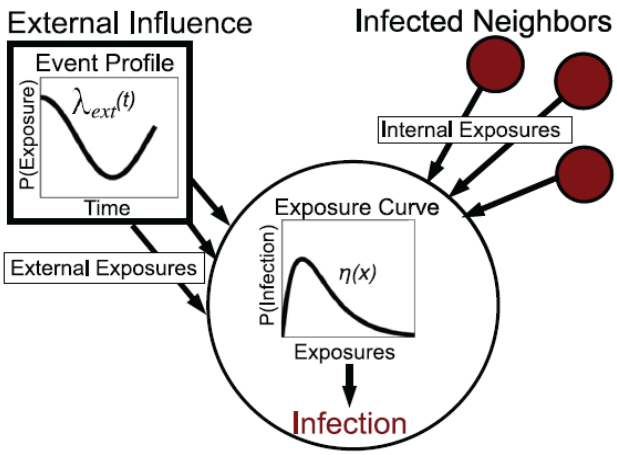
Source: Tang et al. (2009)

2.3.2 Information propagation

This topic originates from the diffusion of innovations (Rogers, 1995). Conceptually, each actor in a network is considered to be either *active* (infected, influenced) or *inactive*, and active node can then spread the information along the edges of the underlying network. A rich set of models has been presented to describe the mechanism by which the information spreads from the infected to an uninfected actors. Among these models, two probability models called independent cascade model (Goldenberg et al., 2001) and linear threshold model (Granovetter, 1978) are most famous. Recently, some interesting works appear in this subfield. Zhao et al. (2011) reveal the importance role of weak-tie actors in the information diffusion process. Myers et al. (2012) consider the impact of external sources on information propagation.

Figure 6 depicts the model in Myers et al. (2012), which distinguishes between *exposures* and *infections*. An exposure event occurs when a node gets exposed to information *I*, and an infection event occurs when a node posts a tweet with information *I*. Exposures to information lead to an infection. A node can get exposed to information in two different ways. First, a node *U* becomes aware of information *I* whenever one of his neighbours in the social network posts a tweet containing *I* (i.e., traditional internal exposure). Second, *U* can be exposed to *I* through the activity of the external source (i.e., external exposure).

Figure 6 The information diffusion model integrating external influence (see online version for colours)



Source: Myers et al. (2012)

2.3.3 Influence maximisation

It aims to find a set of *K* influential actors such that the expected number of actors reached by influence spreading from the selected actor set is maximised. A widely-used motivating application of influence maximisation is *viral marketing* which targets at selecting a small number of influential users to adopt a product, and subsequently trigger a large cascade of further adoptions by utilising the 'word-of-mouth' effect in social

networks (Mahajan et al., 1990). Kempe et al. (2003) formulate the problem as a discrete optimisation problem which is widely adopted by subsequent studies. Since calculating influence spread induced by a given seed sets is very difficult, many heuristics such as degree discount, PMIA, MIAM and MIAC, are presented to ease this problem (Chen et al., 2009, 2010; Liu et al.,(2012).

2.3.4 Summary

There have been and will be an increasing number of research study and practical applications in social influence analysis. Due to the limited space, we are impossible to cover all studies, and we hope that the cited book chapter can point out the missing works.

2.4 Other emerging issues

In this section, we introduce two emerging issues including online advertising and social spam detection. There is few literature about both topics, and therefore this implies that we are faced a lot of opportunities.

2.4.1 Online advertising in social network

In traditional web research area, online advertising aims to match ads with web page content, and thus ads are displayed on the web pages returned by search engine (Lacerda et al., 2006). However, under SN environment, the advertisement problem might change slightly. Provost et al. (2012) consider online brand advertising which focuses on getting a brand-oriented message to an audience of interest. Li et al. (2012) formulate online advertising as viral marketing; that is, given a fixed advertisement investment, e.g., a number of free samples that can be given away to a small number of users, a company needs to determine the probability that users will eventually purchase the product. Besides content (or user) relevance analysis, social influence analysis is often leveraged for online advertising in social network (Bao and Chang, 2010; Li et al., 2012).

2.4.2 Social spam detection

Online social networks are increasingly becoming a source for spreading malware and phishing attacks. Jagatic et al. (2007) found that phishing messages sent by social friends achieved 72%success (measured by clicking on a phishing link in the message). As social spam continues threatening the trusted online social environment, this darker side of the SN has attracted much more attentions recently. At the beginning, numerous studies focused on defending attacks against tagging systems (Koutrika et al., 2008; Ma et al., 2009). Lee et al. (2010) deployed social honeypots to collect evidence of spam behaviour and thus trained a classifier for detection. Recently, more and more attention have been paid to the spammers in Twitter. For instance, spam URLs in tweets were investigated in Grier et al. (2010), the behaviour of spammers and their supporters were analysed in Yang et al. (2012).

2.5 Open issues and trends

Having reviewed the research directions and their state-of-the-art we can conclude that despite recent advances in SN, apart from the aforementioned two emerging areas, there are a number of open issues that need serious and immediate attention. These include:

- *Large-scale*. Social networks continue to evolve and increase in size. Discovering communities, predicting future links, conducting temporal analysis and constructing specific social graphs on large-scale social networks will continue to be a dynamic research challenge.
- *Rich media*. Besides the large-scale, the categories of social media contains blogs, tweets, photos and videos. Moreover, different users have different needs when it comes to the consumption of social media. To organise, index, and retrieve these large-scale rich-media data becomes a big challenge.
- *Personalisation*. Methods and systems for personalisation have potential to improve social interaction and enhance social inclusion. Towards more intelligent SN systems, personalisation models and algorithms have to be studied in a greater extent.
- *Heterogeneity*. Most of the work focus on the analysis of the friendship graph. However, analysis on heterogeneous social graphs opens great opportunities for mining patterns or making predictions on the social graph with multi-kinds of nodes and different relationships.
- *Ethical issues*. Online social communities face also critical social and ethical issues that need special care and delicate handling. Sharing of personal information, protection of child exploitation and many other problems have to be studied and answered appropriately.
- *Business applications*. The novel collaboration paradigms provided by SN is useful to business applications such as customer support, targeted advertising, targeted marketing, and external communications. How to apply SN for these broad applications is a prominent challenge.
- *Mo-Lo-So networks*. The new breed of mobile-friendly, location-aware social (Mo-Lo-So) networks have added new dimensions into SN analysis. Although a wealth of research efforts have been devoted to this area, the analysis for Mo-Lo-So networks is still needed for serious study.
- *Social gaming*. Research is needed on better mass feedback mechanisms for both social gaming and social television. For social gaming as ‘serious game’ is a research challenge.

3 Integrating SN into RSs

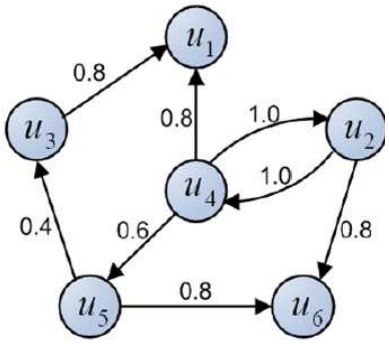
As discussed above, the main stream of RSs is to utilise collaborative filtering, i.e., relying on ratings of like-minded users to predict the preference for the target user via an aggregated way. Under this scheme, the key point is to find the specific user group who share the close taste or interest to the target, where it is done by measure the user interest

similarity based on the user rating vector. This step brings in some new challenges that suffer almost all RS, that is, the problems of cold-start for new user and item, and the sparsity of user rating data. The cold-start problem is caused by the lack of rating by new user or on new items, therefore resulting in the failure of similarity calculation. In contrast to cold-start, the data sparsity seems not too fatal but still greatly degrades the recommendation performance. In another word, without the sufficient user rating data, the recommendation accuracy cannot be guaranteed satisfactorily.

To address these challenges, a lot of research practices have been conducted by integrating complementary information, such as SN of users, item category info, social tags and so on. The key idea of such approaches is to enrich the interactions between users and items. As such, a new type of RS has emerged, namely *social recommendation*. In below section, we review several typical approaches of using SN into RSs. According to where the SN is incorporated and which kinds of recommendation models are used, we categorise them into social-enhanced RS, social regularisation RS, and social-circle RS type.

As discussed the motivation here is to tackle the cold-start and sparsity problems encountered in most RS and make more use of preferences made by her/his close friends. The below mentioned algorithms are designed from the above perspectives. Figure 7(a) depicts an example of social network, which is used to enrich the sparse user-item rating data. In this example, u_4 is a new user, who does not give feedbacks on any items and we aim to make recommendations for this user via referring to his close friends, i.e., u_1, u_2 and u_5.

Figure 7 Illustration of social network and user rating data, (a) social network graph (b) user rating (user-item matrix) (see online version for colours)



(a)

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8
u_1	5	2		3		4		
u_2	4	3			5			
u_3	4		2				2	4
u_4								
u_5	5	1	2		4	3		
u_6	4	3		2	4		3	5

(b)

Source: Ma et al. (2008)

3.1 Using SN as an additional data

One work on integrating SN as an additional data was proposed by Ma et al. (2008) named SoRec. As shown in Figure 7, two matrices are constructed based on the SN of users and the user-item rating data, where the former matrix is used to complement the latter matrix in MF. More specifically, SoRec factorise the SN graph and user-item matrix simultaneously, resulting in U^TZ and U^TV, where the shared low-dimensional matrix U denotes the user latent factor space, Z is the factor space is SN graph, and V

represents the low-dimensional item latent factor space. After the factorisation of matrices, it re-construct the predicted user-item matrix by multiplying U and V. The entry r_{ij} in the product of U^TV indicates the prediction probability of the item i_j preferred by the user u_i. More importantly, for the new user [e.g., u_4 in Figure 7(b)], its ratings on various items are predicted accordingly, reflecting the solving of cold-start problem. The underlying motivation here is that relying on its close friends, the new user's rating is predicted by incorporating it into the full MF. Similar work has been reported in Ma et al. (2009) and Yu et al. (2011).

3.2 Social regulation in MF

Known from above discussion, CF is a dominant approach in most RS and MF has been proven to be more effective and robust that neighbourhood-based approaches. The key point in MF is to decompose the user-item rating matrix to form two latent factor matrixes U and V. As the original user-item matrix is very sparse, in order to avoid the over-fitting problem, a regularisation component is usually engaged to control the influence of U and V in optimising the loss function. Similar to the above scenario, social regularisation-based approaches share the idea of incorporating the social network graph as a optimisation constraint in MF, i.e., focusing on offsetting the loss function in MF.

Generally, the low-dimensional MF can be expressed by:

$$\hat{\mathbf{R}} \approx U^T V$$

where U and V denote the low-dimensional user factor space and item space, respectively. Thus, the goal of MF is to minimise the loss function between the predictions and real observations, which is defined as

$$l = \frac{1}{2}\|R - U^T V\|_F^2$$

where R is the observed ratings, which contain a large number of missing values. Since we only rely on the observed ratings in matrix R to conduct factorisation, it is unavoidable to incur in the over-fitting problem, which is very common in machine learning. To overcome this, a regularisation part is introduced into the optimisation.

$$\min_{U,V} \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}\left(R_{ij} - U_i^T V_j\right)^2 + \frac{\lambda_1}{2}\|U\|_F^2 + \frac{\lambda_2}{2}\|V\|_F^2$$

where $\frac{\lambda_1}{2}\|U\|_F^2|\frac{\lambda_2}{2}\|V\|_F^2$ is the regularisation part.

The motivation behind social regularisation is the hypothesis that sometimes, in order to make a decision, we'd like to consult lots of our friends for valuable suggestions. This assumption is often held in real applications and becomes more realistic and practical with the prevalence of SN. In Ma et al. (2011), two algorithms have been proposed, from the perspectives of average and individual regularisation. The former is to add the averaged ratings from friends as a constraint to guide the learning, while the latter is to leverage the individual rating into regularisation.

- Average-based regularisation

$$\begin{aligned}&\min_{U,Vm} L_1(R,U,V)=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}\left(R_{ij}-U_i^T V_j\right)^2\\&+\frac{\alpha}{2}\sum_{i=1}^{m}\left\|U_i-\frac{1}{F+(i)}\sum_{f\in F+(i)} U_f\right\|_F^2+\frac{\lambda_1}{2}\|U\|_F^2+\frac{\lambda_2}{2}\|V\|_F^2\end{aligned} \tag{18}$$

- Individual-based regularisation

$$\begin{aligned}&\min_{U,V} L_1(R,U,V)=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}\left(R_{ij}-U_i^T V_j\right)^2\\&+\frac{\beta}{2}\sum_{i=1}^{m}\sum_{f\in F+(i)} sim(i,f)\left\|U_i-U_f\right\|_F^2+\frac{\lambda_1}{2}\|U\|_F^2+\frac{\lambda_2}{2}\|V\|_F^2\end{aligned} \tag{19}$$

3.3 Friend-trust-based social RS

Recently with the influx of SN, trust relations between friends are becoming an important factor in RSs. The key idea here is to extend social relation to social trust concept, where users are often referring to their friends' rating for recommendations. Jamali and Ester (2009) exploited the trust information to reinforce the item-based recommendation via a random walk manner. The random walk model provided a way to define and to measure the confidence of a recommendation. In Jamali and Ester (2010), a new social recommendation algorithm based on social MF was proposed, named SocialMF. The hypothesis underlying SocialMF is that neighbours in the social network may have similar interest. This similarity is enforced in the regularisation part in addition to the regularisation part of U and V, implying that the user profile Q_u should be similar to the (weighted) average of his/her friends' profiles Q_v (measured in terms of the squared error):

$$\begin{aligned}&\frac{1}{2}\sum_{(u,i)obs}\left(R_{u,i}-\hat{R}_{u,i}\right)^2\\&+\frac{\beta}{2}\sum_{allu}\left(\left(Q_u-\sum_v S_{u,v}^*\right)\left(Q_u-\sum_v S_{u,v}^*\right)^T\right)+\frac{\lambda}{2}\left(\|P\|_F^2+\|Q\|_F^2\right)\end{aligned} \tag{20}$$

where the weighted social relationship of user u with user v (e.g., user u trusts/knows/follows user v) is represented by a positive value $S_{u,v}^* \in (0,1]$.

Many online social networks now support a new feature of Friend-Circles. A user trusts different subsets of friends in different domains. For example, a user u may trust user v in *Cars* category while not trust v in *Kids TV Show* category. How to differentiate the various trust circles in different categories/topics to predict ratings is becoming a new challenge. Yang et al. (2012) proposed a circle-based social recommendation algorithm, namely *CircleCon* by leveraging the circle information embedded in social networks, e.g., circles in *Google+* or *Facebook*. Different from SociaMF, which enforces the overall interest difference in regularisation, CircleCon calculates the interest difference on a basis

of category individually, and then accumulate the total difference measured in terms of squared error. Thus in this situation, the optimisation problem is re-written as:

$$\begin{aligned}\ell^{(c)}\left(R^{(c)}, Q^{(c)}, P^{(c)}, S^{(c)*}\right) &= \frac{1}{2}\sum_{(u,i)obs}\left(R_{u,i}-\hat{R}_{u,i}\right)^2 \\ &+\frac{\beta}{2}\sum_{allu}\left(\left(Q_u^{(c)}-\sum_v S_{u,v}^{(c)*}Q_v^{(c)}\right)\left(Q_u-\sum_v S_{u,v}^{(c)*}Q_v^{(c)}\right)^T\right) \\ &+\frac{\lambda}{2}\left(\left\|P^{(c)}\right\|_F^2+\left\|Q^{(c)}\right\|_F^2\right)\end{aligned} \tag{21}$$

Unfortunately, in most existing multi-category rating datasets, a user's social connections from all categories are mixed together. So if we use all social trust information for rating prediction in a specific category, we misuse social trust information from other categories, which compromises the rating prediction accuracy. Apart from that, even if the circles were explicitly known, e.g., Circles in Google+ or Facebook, they may not correspond to particular item categories that a RS may be concerned with. Therefore, inferred circles concerning each item-category may be of value by themselves, besides the explicitly known circles.

4 Social tagging RSs

In past years, the emergence of Web 2.0 applications has created a new era for sharing and organising resources in online social communities. The shared resources could range diversely from the social bookmarks *de.licio.us*[3] to medical articles in *MedWorm*[4] and scientific publications on *CiteULike*[5]. These we sites have one concept in common: the phenomenon of *folksonomy*-users choose free style terms (i.e., tags) to annotate various resources indicating their own perceptions or conceptual judgments on these resources for better indexing and annotation. In other words, *Tag*, as one kind of specific lexical information that is user-generated metadata with uncontrolled vocabulary, plays a crucial role in such social collaborative tagging systems. In addition to keywords or terms contained in the resources, tagging provides a complementary feature for web resource and can for example reveal user objective opinions or comments, which could be used for information searching and retrieval. Tags apparently convey the semantic conceptual information on resources collaboratively generated by users to some extents. Utilising tag information, therefore, could undoubtedly foster search capabilities in social tagging systems. Recently tagging has been widely used in RSs for many applications (Durao and Dolog, 2010; Jäschke et al., 2007; Tso-Sutter et al., 2008).

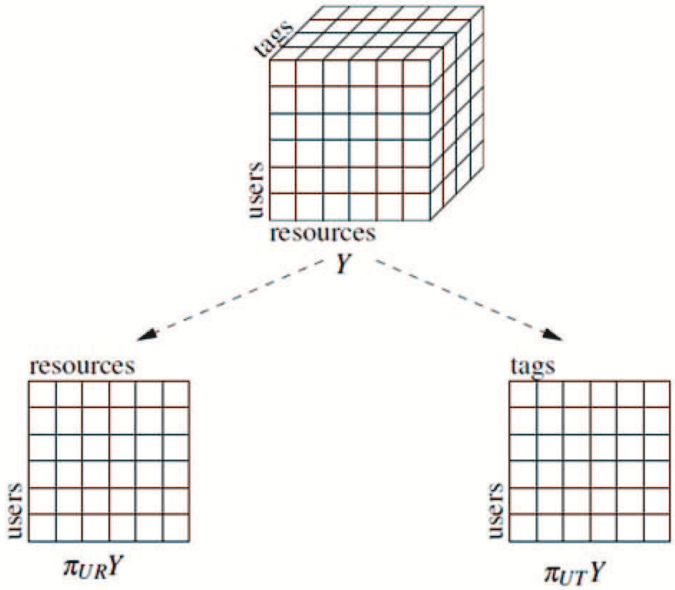
Despite of the considerable successes achieved in tag-based RSs and the huge column of tag data available in internet, the most significant burdens that such kinds of systems facing are the problems of ambiguity, redundancy and less semantics of tags, the incapability of users issuing appropriate tags and the sparsity of tagging data due to intrinsic characteristics of social tagging behaviour. For example, due to the nature of folksonomy, even for a well known web page like *Microsoft's* homepage, it is very likely that different users choose their own words as tags, such as an abbreviation term *OS* or a polysemy word *Window* or even a non-existing word caused by spelling mistakes. Moreover, the users sometimes may do not have sufficient domain knowledge on a

specific topic thus they could not choose appropriate tags or only annotate a small fraction of resources resulting in the serious problem of sparsity in tag data (Rendle et al., 2009). As discussed above, due to the problem of less semantics and sparse quality of tag data, the tag-based RSs which rely simply on the lexical similarity of tags alone are likely to neglect the retrieval of some closely related resources. As a result, the lack of satisfactory quality of metadata largely impairs the retrieval of needed resources since the resources without appropriate annotations are obviously more difficult to retrieve than those properly tagged resources (Budura et al., 2009).

4.1 Social tagging data model

The conventional data structure of tagging data is expressed as three-dimensional array, which is shown in Figure 8. A typical social tagging system has three types of entities, users, tags and resources which are interrelated with one another. Social tagging data can be viewed as a set of triples (Heymann et al., 2008; Guan et al., 2010). Each triple (u, t, r) represents an observation of a user u annotating a tag t on a resource r. A social tagging system can be described as a four-tuple collection – there exist a set of users, U; a set of tags, T; a set of resources, R; and a set of annotations, A^n. We denote the data in the social tagging system as D and define it as: $D = <U, T, R, A^n>$. The annotations, A^n, are represented as a set of triples containing a user, tag and resource defined as: $A^n \subseteq <u, t, r>$: $u \in U$, $t \in T$, $r \in R$. Therefore, a social tagging system can be viewed as a tripartite hypergraph (Mika, 2005) with users, tags and resources represented as nodes and the annotations represented as hyper-edges connecting users, tags and resources. As this data model reflects the interactions between folks, terms and documents, it is often call folksonomy model (Hotho et al., 2006).

Figure 8 Folksonomy data model



4.2 Standard tag-based recommendation

The standard tag-based recommendation is principally similar to a process of traditional information retrieval but with an additional input of the user tagging preference for personalisation (or called personalised recommendation). The procedure consists of two steps of search and personalisation. The first step produces a list of candidate resources r_s based on the similarity computation between the query tag issued by a user and all resources in terms of term frequency – inverse document frequency (tf-idf).

The second step utilises the tagging preference of users to make the personalisation. Under the vector space model, each user, u, is modelled as a vector (also called user profile) over a set of tags, where $w(t_i)$, in each dimension corresponds to the relationship of a tag t_i with this user, u, $\vec{u} =< w(t_1), w(t_2), ..., w(t_{|T|}) >$. Likewise each resource, r, can be modelled as a vector (i.e., resource profile) over the same set of tags, $\vec{r} =< v(t_1), v(t_2), ..., v(t_{|T|}) >$, where $v(t_i)$ represents the relationship of a tag t_i with this resource. After that, the similarity computation, e.g., cosine measure, of the target user profile u and the candidate resource profiles r_s selected by the first step, is performed, $sim(u, r)$, $r \in r_s$, to further generate the personalised resources based on various recommendation strategies. The distinction of the tag-based recommendation from the standard information search is that here the recommendation is derived from, not only the query itself, but also the user tagging preference (i.e., personalisation). Obviously, the straight advantage of personalised recommendation is able to provide users more personalised and preferable information by leveraging the additional metadata, i.e., tag.

Many researches have been done in this area to leverage tags for personalised recommendations. Durao and Dolog (2010) developed a multi-factorial tag-based RS, which took various lexical and social factors of tags into the similarity calculation. Shepitsen et al. (2008) proposed a personalised recommendation system by using hierarchical clustering. In this approach, instead of using the pure tag vector expressions, a preprocessing on tag clustering was performed to find out the tag aggregates for personalised recommendation. Zhang et al. (2010) aimed to integrate the diffusion on user-item-tag tripartite graphs to improve the recommendation of state-of-the-art techniques.

4.3 State-of-the-art in social tag-based RS

4.3.1 Multi-mode recommendations

Different from tradition RS, where the goal is for rating predictions or resources recommendations users concern, social tag-based RS (STS) provides a flexible way for multi-mode recommendations, i.e., user, item or tag recommendation.

Item recommendation is the most commonly used recommendation paradigm, which is to recommend potentially interested items to users based on the user tagging behaviours. Related studies include the work in Guan et al. (2010), which modelled the triadic relations with a graph and designed a multi-type interrelated objects embedding (MIOE) algorithm to find documents. Xu et al. (2011) investigated the use of semantic enhancement in STS for improved item recommendations. The proposed technique combined clustering and latent topic model in a fusion way to capture the similarity between users and items explicitly and implicitly.

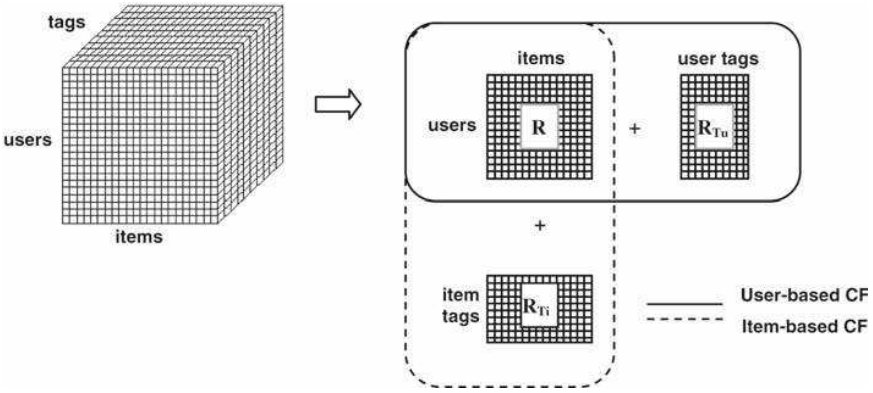
Tag recommendation is also widely studied in STS, which aims to recommend tags possibly interested by users. Krestel et al. (2009) proposed an algorithm to reveal the hidden topic structure embedded in social tagging data for recommending tags for users. Guan et al. (2009) devised a tag recommendation algorithm based on the node ranking in graph and experimental study demonstrated the superiority of the proposal. Yang et al. (2012) developed an innovative algorithm for tag recommendation by using rank walk with restart with supervision of annotated tags. The main contribution of this paper is the adoption of supervised random walk.

User recommendation is not a very popular application in STS but useful in human focused applications such as SN. For this task, the aim is to suggest a list of users to the target user, who may form the neighbourhood. Li et al. (2008) reported an interesting study on discovering user common interest by calculating user similarity over items. Besides this, many work have been SNA, such as link prediction described in the previous section. For example, Tang et al. (2012) proposed a cross-domain topic learning (CTL) algorithm for recommending cross-domain collaborator.

4.3.2 Machine learning deployment in STS

As mentioned above, STS heavily relies on the multi-aspect tagging behaviours, which opens a big room for machine learning algorithms to be employed. The multi-dimensional nature of social tagging data raises new demands for more advanced techniques, that apparently intuitive approach e.g., user-/item-based CF, cannot handle. One kind of representative methods are to deal with the triadic relations amongst users, items and tags via high dimensional mathematical models. Symeonidis et al. (2010) proposed a unified framework *HOSVD* for ternary semantic analysis based on the loss function optimisation. With the decomposed ternary expressions, the multi-mode recommendations can be implemented in terms of users, items and tags. This approach systematically explored the utilisation of ternary semantic analysis (i.e., tensor decomposition) in social tagging systems. In contrast, Rendle et al. (2009) devised an effective loss function optimisation algorithm based on maximising the objective function of area under curve (AUC). Different from other tensor factorisation algorithms, his approach was to construct the latent factorisation based on the AUC optimisation, whose idea has inspired the algorithm of supervised random walk.

Figure 9 Extended CF model with tags



Another essential work direction is to extend the traditional CF by expanding tag as an additional aspect of user or item, resulting in an expanded user-/item-based CF. The typical work was published in Tso-Sutter et al. (2008), whose key idea was shown in Figure 9. Likewise, Hofmann (2004) employed probabilistic latent semantic analysis (PLSA) on the projected tagging data (i.e., user-tag or item-tag matrix) to capture the similarity of users and items.

4.3.3 Fusion of social relations

Recently, with the increasing development of social websites and appearance of social data, researchers have begun to pay attention to the social data and explored its usage in RSs. Groh G. (2007) used social network data for neighbourhood generation. Konstans et al. (2009) adopted random walk with restart to model the social tagging in a music track recommendation system. In addition, Hummel et al. (2007) proposed an online social RS attempting to use more social information for recommendation generation. All the work shows that their fusion social information can benefit the RS. However, their work mainly focuses on friendship, i.e., the similarity between users'. Compared to friendship, the community relationships, or called group, are used in Spertus et al. (2005) and Chen et al. (2009) for recommendations.

4.3.4 Open issues and trends

Although researches in RS has progressed significantly with the supports from related research domains, such as machine learning, information retrieval and social and behavioural science, as well as the attention and investment from industries, there is still a huge room of research and application issues needed to be tackled by joint efforts from academia and industry. These include:

- *Prediction or ranking*. The traditional RS focuses more on predicting ratings for users accurately via collective preference. However, due to the lack of sufficient user ratings, the satisfactory prediction is not an easy job. Recently, researchers gradually realised that the ranking, in fact, is also an essential issue in RS. With the empowerment from search engine community, recommendation is becoming easy to implement.
- *Algorithms or applications*. From the aim and domain perspective, RS is indeed an application-oriented technique rather than the algorithmic development. The real requirement from industry and business will stimulate the advance of technologies and the emergence of new services. Establishing close contacts with industry is an unavoidable trend in this area.
- *Separate or ensemble*. The researches and practices in RS are not isolated from the connection to related research efforts in machine learning and data mining. In fact, RS are an ensemble of algorithms and techniques of related disciplines. Specially, with the emphasis on multi-dimensional and heterogeneous big data, machine learning and data mining play a key role in address real problems. As a RS designer and developer, we need not only the skills about system implementation, but also the knowledge about theories and algorithms. The broad and tight connections to core theoretical research communities are widely recognised by us.

- *Big data issue*. Big data is becoming an unprecedented concern and focus due to the large availability of huge and heterogeneous data provisions. In RS, exploiting auxiliary and complementary data from various sources becomes a common trend in improving recommendation performance. Due to the complicated structure of data, how to acquire, integrate, consolidate and make use of them, and where the cutting edge techniques would be employed have attracted people's attention and interest. It is undoubted that RS is one part of issues involved in big data acquisition, analysis, and utilisation.
- *RS with mobile networking*. With the prevalence of pervasive mobile computing and social Web, mobile networking is becoming an indispensable part of our daily life. The use of mobile creates a convenience for users to access information over the Internet and exchange thoughts and opinions with others in a 7/24 way. It is expected that mobile-based recommendation services, such as location-based SN evidences a new push of technologies.

5 Conclusions

RSs are softwares and systems predicting preferences of users on various items that are of interest to users. The suggestions provided are aimed at improving user experience and loyalty, facilitating decision-making for users and creating more revenues for online businesses and merchants and so on. With the prevalence of web applications, web-based RS has been widely researched, developed and deployed in a variety of real applications and commercial systems. Development of RSs is a multi-disciplinary effort which involves experts from various fields such as artificial intelligence, human computer interaction, information technology, data mining, statistics, adaptive user interfaces, decision support systems, marketing, or consumer behaviour. RS has evolved for a certain period, starting from the simple demography-based RS, content-based RS to the predominance of collaborative filtering techniques. The performance of RS is heavily dependent on the user historical feedbacks and ratings, while mostly they are very sparse or do not exist at all, e.g., for new users and items. Incorporating complementary information into recommendations is becoming a technical trend and common ways in not only research academia but also application domains.

SN is an online service, platform, or site that focuses on facilitating the building of social networks or social relations among people who, for example, share interests, activities, backgrounds, or real-life connections. SN provides the valuable information in terms of user social relations, which in turn, assists in complementing the social aspects in RS. Although SNA is originated from social and behavioural research, nowadays it is becoming an active and influential topic in theoretical and industrial practices with thanks to the empowerment of computational and networking capability, evidenced by a large spectrum of SN services, such as *Facebook*, *Google+*, and *Twitter*.

The marriage of RS and SN initiates the emergence of the promising applications in RS – social RS. One distinctive advantage of social RS over traditional RS is the capability of tacking sparsity and cold-start problems suffering most RS. This survey has systematically reviewed the state-of-the-art techniques and algorithms from the perspectives of RS and SN themselves, as well as the latest advance in integrating the SN in recommendations, presenting the landscape of researches and applications conducted

in related areas and outlining some potential future research directions and open questions.

Acknowledgements

This research was partially supported by National Natural Science Foundation of China under Grants 61103229 and 71372188, National Center for International Joint Research on E-Business Information Processing under Grant 2013B01035, National Key Technologies R&D Program of China under Grants 2013BAH16F01 and 2013BAH16F04, Industry Projects in Jiangsu S&T Pillar Program under Grant BE2012185, Key Project of Natural Science Research in Jiangsu Provincial Colleges and Universities under Grant 12KJA520001, and the National Soft Science Research Program under Grant 2013GXS4B081.

References

Abello, J., Resende, M.C. and Sudarsky, S. (2002) 'Massive quasi-clique detection', *LATIN'02*.

Agarwal, R. and Srikant, R. (1994) 'Fast algorithms for mining association rules in large databases', *VLDB'94*, pp.487–499.

Agarwal, R. and Srikant, R. (1995) 'Mining sequential patterns', *ICDE'95*, pp.3–14.

Agarwal, R., Aggarwal, C. and Prasad, V. (1999) 'A tree projection algorithm for generation of frequent itemsets', *Journal of Parallel and Distributed Computing*, Vol. 61, No. 3, pp.350–371

Backstrom, L., Sun, E. and Marlow, C. (2010) 'Find me if you can: improving geographical prediction with social and spatial proximity', *WWW'10*.

Bao, H. and Chang, E. (2010) 'Adheat: an influence-based diffusion model for propagating hints to match ads', *WWW'10*.

Borgatti, S.P. and Everett, M.G. (1999) 'Models of core periphery structures', *Social Networks*, Vol. 21, No. 4, pp.375–395.

Boykin, S. and Merlino, A. (2000) 'Machine learning of event segmentation for news on demand', *Communication of the ACM*, Vol. 43, No. 2, pp.35–41.

Büchner, A.G. and Maurice, D.M. (1998) 'Discovering internet marketing intelligence through online analytical web usage mining', *SIGMOD Record*, Vol. 27, No. 4, pp.54–61.

Budura, A., Michel, S., Cudré-Mauroux, P. and Aberer, K. (2009) 'Neighborhood-based tag prediction', *ESWC'09 on The Semantic Web*, pp.608–622.

Chakrabarti, S. (2000) 'Datamining for hypertext: a tutorial survey', *SIGKDD Exploration Newsletter*, Vol. 1, No. 11, pp.1–11.

Chen, C. and Chen, M. (2012) 'Tscan: a content anatomy approach to temporal topic summarization', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No. 1, pp.170–183.

Chen, J., Geyer, W., Muller, M.J. and Guy, I. (2009a) 'Make new friends, but keep the old: recommending people on social networking sites', *ChI'09*, pp.201–210.

Chen, W., Wang, Y. and Yang, S. (2009b) 'Efficient influence maximization in social networks', *KDD'09*, pp.199–208.

Chen, W., Wang, C. and Wang, Y. (2010) 'Scalable influence maximization for prevalent viral marketing in large-scale social networks', *KDD'10*, pp.1029–1038.

Cho, E., Myers, S.A. and Leskovec, J. (2011) 'Friendship and mobility: user movement in location-based social networks', *KDD'11*, pp.1082–1090.

Clauset, A. (2005) 'Finding local community structure in networks', *Physical Review E*, Vol. 72, No. 2, p.026132.

Clauset, A., Moore, C. and Newman, M.E.J. (2008) 'Finding local community structure in networks', *Physical Review E*, Vol. 453, No. 7191, pp.98–101.

Cohen, E., Krishnamurthy, B. and Rexford, J. (1998) 'Improving end-to-end performance of the web using server volumes and proxy items', *SIGCOMM'98*, pp.241–253.

Cranshaw, J., Toch, E., Hong, J.I., Kittur, A. and Sadeh, M. (2010) 'Bridging the gap between physical location and online social networks', *UbiComp*, pp.119–128.

Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K. and Slattery, S. (1998) 'Learning to extract symbolic knowledge from the World Wide Web', *AAAI'98*, pp.509–516.

Culotta, A. (2010) 'Learning to extract symbolic knowledge from the World Wide Web', CoRR, abs/1007.4748.

David, J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D. and Kleinberg, J. (2010) 'Inferring social ties from geographic coincidences', *Proceedings of the National Academy of Sciences (PNAS)*, Vol. 107, No. 52, pp.22436–22441.

Doppa, J.R., Yu, J., Tadepalli, P. and Getoor, L. (2009) 'Chance-constrained programs for link prediction', *Workshop on Analyzing Networks and Learning with Graphs at NIPS'09*.

Dunja, M. (1996) *Personal Web Watcher: Design and Implementation (Report)*, Technical Report IJS-DP-7472.

Durao, F. and Dolog, P. (2010) 'Extending a hybrid tag-based recommender system with personalization', *SAC'10*, pp.1723–1727.

Eagle, N., Pentland, A. and Lazer, D. (2010) 'Inferring friendship network structure by using mobile phone data', *Proceedings of the National Academy of Sciences (PNAS)*, Vol. 106, No. 36, pp.15274–15278.

Eubank, S., Guclu, H., Kumar, V.S.A., Marathe, M.V., Srinivasan, A., Toroczkai, T. and Wang, N. (2004) 'Modelling disease outbreaks in realistic urban social networks', *Nature*, Vol. 429, No. 6988, pp.180–184.

Fortunato, S. (2010) 'Community detection in graphs', *Physics Reports*, Vol. 486, Nos. 3–5, pp.75–174.

Ge, Y., Xiong, H., Tuzhilin, A., Gruteser, M. and Pazzani, M. (2010) 'An energy-efficient mobile recommender system', *KDD'10*, pp.899–908.

Ghani, R. and Fano, A. (2002) 'Building recommender systems using a knowledge base of product semantics', *Workshop on Recommendation and Personalization in E-Commerce at AHAWS'02*, pp.899–908.

Glaeser, E. and Kahn, M. (2004) 'Sprawl and urban growth', *Handbook of Regional and Urban Economics*, Chapter 56, Vol. 4, pp.2481–2527.

Goldenberg, J., Libai, B. and Muller, E. (2001) 'Talk of the network: a complex systems look at the underlying process of word-of-mouth', *Marketing Letters*, Vol. 12, No. 3, pp.211–223.

Granovetter, M. (1978) 'Threshold models of collective behavior', *American Journal of Sociology*, No. 6, pp.1420–1443.

Granovetter, M.S. (1973) 'The strength of weak ties', *American Journal of Sociology*, Vol. 78, No. 6, pp.1360–1380.

Grier, C., Thomas, K., Paxson, V. and Zhang, M. (2010) 'spam: the understanding on 140 characters or less', *CCS'10*, pp.27–37.

Groh, G. (2007) 'Recommendations in taste related domains: collaborative filtering vs. social filtering', *Group'07*, pp.127–136.

Guan, Z., Bu, J., Mei, Q., Chen, C. and Wang, C. (2009) 'Personalized tag recommendation using graph-based ranking on multi-type interrelated objects', *SIGIR'09*, pp.540–547.

Guan, Z., Wang, C., Bu, J., Chen, C., Yang, K., Cai, D. and He, X. (2010) 'Document recommendation in social tagging services', *WWW'10*, pp.391–400.

Han, E., Karypis, G., Kumar, V. and Mobasher, B. (1998) 'Hypergraph based clustering in high dimensional data sets: a summary of results', *IEEE Data Engineering Bulletin*, Vol. 21, No. 1, pp.15–22.

Handcock, M.S., Raftery, A.E. and Tantrum, J.M. (2007) 'Model-based clustering for social networks', *Journal of the Royal Statistical Society Series A*, Vol. 127, No. 2, pp.301–354.

Hasan, M.A. and Zaki, M. (2011) 'A survey of link prediction in social networks', *Social Network Data Analytics*, Vol. 127, No. 2, pp.243–275.

Hasan, M.A., Chaoji, V., Salem, S. and Zaki, M. (2006) 'Link prediction using supervised learning', *Workshop on Link Analysis, Counterterrorism, and Security at SDM'06*.

Herlocker, J., Konstan, J., Borchers, A. and Riedl, J. (1999) 'An algorithmic framework for performing collaborative filtering', *SIGIR'99*, pp.230–237.

Herlocker, J., Konstan, J., Terveen, L. and Riedl, J. (2004) 'Evaluating collaborative filtering recommender systems', *ACM Transaction on Information Systems*, Vol. 22, No. 1, pp.5–53.

Heymann, P., Ramage, D. and Garcia-Molina, H. (2008) 'Social tag prediction', *SIGIR'04*, pp.531–538.

Hofmann, T. (2004) 'Latent semantic models for collaborative filtering', *ACM Transactions on Information Systems*, Vol. 22, No. 1, pp.89–115.

Hotho, A., Jäschke, R., Schmitz, C. and Stumme, G. (2006) 'Folkrank: a ranking algorithm for folksonomies', *FGIR'06*, pp.111–114.

Hou, J. and Zhang, Y. (2003) 'Effectively finding relevant web pages from linkage information', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15, No. 4, pp.940–951.

Hummel, H., Berg, B., Berlanga, A., Drachsler, J.J. Nadolski, R. and Koper, R. (2007) 'Combining social-based and information-based approaches for personalized recommendation on sequencing learning activities', *International Journal of Learning Technology*, Vol. 3, No. 2, pp.152–168.

Jagatic, T.N., Johnson, N.A., Jakobsson, M. and Menczer, F. (2007) 'Social phishing', *Communications of the ACM*, Vol. 50, No. 10, pp.94–100.

Jamali, M. and Ester, M. (2009) 'Trustwalker: a random walk model for combining trust-based and item-based recommendation', *KDD'09*, pp.397–406.

Jamali, M. and Ester, M. (2010) 'A matrix factorization technique with trust propagation for recommendation in social networks', *RecSys'10*, pp.135–142.

Jäschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L. and Stumme, G. (2007) 'Tag recommendations in folksonomies', *PKDD'07*, pp.506–514.

Jin, X. and Mobasher, B. (2003) 'Using semantic similarity to enhance item-based collaborative filtering', *IKS'03*.

Joachims, T., Freitag, D. and Mitchell, T. (1997) 'Web watcher: a tour guide for the World Wide Web', *IJCAI'97*, pp.770–777.

Karrer, B. and Newman, M.E.J. (2011) 'Stochastic block models and community structure in networks', *Physical Review E*, Vol. 83, No. 1, p.016107.

Kashima, H., Kato, T., Yamanishi, Y., Sugiyama, M. and Tsuda, K. (2009) 'Link propagation: a fast semi-supervised learning algorithm for link prediction', *SDM'09*, pp.1099–1110.

Kempe, D., Kleinberg, J. and Tardos, É. (2003) 'Maximizing the spread of influence through a social network', *KDD'03*, pp.137–146.

Kernighan, B.W. and Lin, S. (1970) 'An efficient heuristic procedure for partitioning graphs', *Bell Systems Technical Journal*, Vol. 49, No. 1, pp.291–307.

Kleinberg, J.M. (1998) 'Authoritative sources in a hyperlinked environment', *SODA'98*, pp.668–677.

Kleinberg, J.M. (2002) 'Bursty and hierarchical structure in streams', *KDD'02*, pp.91–101.

Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L. and Riedl, J. (1997) 'Grouplens: applying collaborative filtering to use net news', *Communications of the ACM*, Vol. 40, No. 3, pp.77–87.

Konstas, I., Stathopoulos, V. and Jose, J.M. (2009) 'On social networks and collaborative recommendation', *SIGIR'09*, pp.195–202.

Konstas, Y. (2008) 'Factorization meets the neighborhood: a multifaceted collaborative filtering model', *KDD'08*, pp.426–434.

Kosala, R. and Blockeel, H. (2000) 'Web mining research: a survey', *SIGKDD Exploration Newsletter*, Vol. 2, No. 1, pp.1–15.

Kossinets, G. and Watts, D.J. (2006) 'Empirical analysis of an evolving social network', *Science*, Vol. 311, No. 5757, pp.88–90.

Koutrika, G., Effendi, F., Gyongyi, Z., Heymann, P. and Garcia-Molina, H. (2008) 'Combating spam in tagging systems: an evaluation', *ACM Transactions on the Web*, Vol. 2, No. 4, pp.1–34

Krestel, R., Fankhauser, P. and Nejdl, W. (2009) 'Latent Dirichlet allocation for tag recommendation', *RecSys'09*, pp.61–68.

Lacerda, A., Cristo, M., Gonçalves, M., Fan, W. and Ribeiro-Neto, B. (2006) 'Learning to advertise', *SIGIR'06*, pp.549–556.

Lee, K., Caverlee, J. and Webb, S. (2010) 'Uncovering social spammers: social honeypots + machine learning', *SIGIR'10*, pp.435–442.

Leroy, V., Cambazoglu, B. and Bonchi, F. (2010) 'Cold start link prediction', *KDD'10*, pp.393–402.

Li, X., Guo, L. and Zhao, Y. (2008) 'Tag-based social interest discovery', *WWW'10*, pp.675–684.

Li, Y., Zhao, B. and Lui, J. (2012) 'On modeling product advertisement in large-scale online social networks', *IEEE/ACM Transactions on Networking*, Vol. 20, No. 5, pp.1412–1425.

Liben-Nowell, D. and Kleinberg, J. (2007) 'The link-prediction problem for social networks', *Journal of the American Society for Information Science and Technology*, Vol. 58, No. 7, pp.1019–1031.

Lieberman, H. (1995) 'Letizia: an agent that assists web browsing', *IJCAI'95*, pp.924–929.

Lin, Y., Chi, Y., Zhu, S., Sundaram, H. and Tseng, B. (2009) 'Analyzing communities and their evolutions in dynamic social networks', *ACM Transactions on Knowledge Discovery from Data*, Article 8, Vol. 3, No. 2.

Liu, B., Cong, G., Xu, D. and Zeng, Y. (2012) 'Time constrained influence maximization in social networks', *ICDM'12*, pp.439–448.

Ma, H., King, I. and Lyu, M. (2009) 'Learning to recommend with social trust ensemble', *SIGIR'09*, pp.203–210.

Ma, H., Yang, H., Lyu, M. and King, I. (2008) 'Sorec: social recommendation using probabilistic matrix factorization', *CIKM'08*, pp.931–940.

Ma, H., Zhou, D., Liu, C., Lyu, M. and King, I. (2011) 'Recommender systems with social regularization', *WSDM'11*, pp.287–296.

Mahajan, V., Muller, E. and Bass, F.M. (1990) 'New product diffusion models in marketing: a review and directions for research', *The Journal of Marketing*, pp.1–26.

Mei, Q. and Zhai, C. (2005) 'Discovering evolutionary theme patterns from text: an exploration of temporal text mining', *KDD'05*, pp.198–207.

Meng, X., Wei, F., Liu, X., Zhou, M., Li, S. and Wang, H. (2012) 'Entity-centric topic-oriented opinion summarization in twitter', *KDD'12*, pp.379–387.

Mika, P. (2005) 'Ontologies are us: a unified model of social networks and semantics', *ISWC'05*, pp.522–536.

Mobasher, B., Cooley, R. and Srivastava, J. (1999) 'Creating adaptive web sites through usage-based clustering of urls', *KDEX'99*, pp.19–25.

Mobasher, B., Dai, H., Nakagawa, M. and Luo, T. (2002) 'Discovery and evaluation of aggregate usage profiles for web personalization', *Data Mining and Knowledge Discovery*, Vol. 6, No. 1, pp.61–82.

Mokken, R. (1979) 'Cliques, clubs and clans', *Quality and Quantity*, Vol. 13, No. 2, pp.161–173.

Myers, S., Zhu, C. and Leskovec, J. (2012) 'Information diffusion and external influence in networks', *KDD'12*, pp.33–41.

Newman, M.E.J. (2004) 'Fast algorithm for detecting community structure in networks', *Physical Review E*, Vol. 69, No. 6, pp.33–41.

O'Conner, M. and Herlocker, J. (1999) 'Clustering items for collaborative filtering', *ACM SIGIR Workshop on Recommender Systems*.

Palla, G., Derenyi, I., Farkas, I. and Vicsek, T. (2005) 'Uncovering the overlapping community structure of complex networks in nature and society', *Nature*, Vol. 435, No. 7043, pp.814–818.

Perkowitz, M. and Etzioni, O. (1998) 'Adaptive web sites: automatically synthesizing web pages', *AAAI'98*, pp.727–732.

Perkowitz, M. and Etzioni, O. (1999) 'Adaptive web sites: conceptual cluster mining', *IJCAI'99*.

Pietilainen, A. and Diot, C. (2012) *Dissemination in Opportunistic Social Networks: The Role of Temporal Communities*, Technical report, Technicolor.

Provost, F., Dalessandro, B., Hook, R., Zhang, X. and Murray, A. (2012) 'Audience selection for on-line brand advertising: privacy-friendly social network targeting', *KDD'09*, pp.707–716.

Qamra, A., Tseng, B. and Chang, E. (2006) 'Mining blog stories using community-based and temporal clustering', *CIKM'06*, pp.58–67.

Raymond, R. and Kashima, H. (2010) 'Fast and scalable algorithms for semi-supervised link prediction on static and dynamic graphs', ECML/*PKDD'10*, pp.131–147.

Rendle, S., Marinho, L., Nanopoulos, A. and Schmidt-Thieme, L. (2009) 'Learning optimal ranking with tensor factorization for tag recommendation', *KDD'09*, pp.727–736.

Rogers, E.M. (1995) *Diffusion of Innovations*, Simon and Schuster, New York.

Sarwar, B., Karypis, G., Konstan, J. and Reidl, J. (2001) 'Item-based collaborative filtering recommendation algorithms', *WWW'01*, pp.285–295.

Scripps, J., Tan, P. and Esfahanian, A. (2009) 'Measuring the effects of preprocessing decisions and network forces in dynamic network analysis', *KDD'09*, pp.747–756.

Serrat, O. (2009) *Social Network Analysis*, Technical report, Cornell University.

Shardanand, U. and Maes, P. (1995) 'Social information filtering: algorithms for automating word of mouth', *CHI'95*, pp.210–217.

Shepitsen, A., Gemmell, J., Mobosher, B. and Burke, R. (2008) 'Personalized recommendation in social tagging systems using hierarchical clustering', *RecSys'08*, pp.259–266.

Siaw, D., Ngu, W. and Wu, X. (1997) 'Sitehelper: a localized agent that helps incremental exploration of the World Wide Web', *Computer Networks*, Vol. 29, Nos. 8–13, pp.1249–1255.

Singla, P. and Richardson, M. (2008) 'Yes, there is a correlation:-from social networks to personal behavior on the web', *WWW'08*, pp.655–664.

Slater, P. (2008) *Established Clustering Procedures for Network Analysis*, Technical Report, arXiv: 0806.4168.

Spertus, E., Sahami, M. and Buyukkokten, O. (2005) 'Evaluating similarity measures: a large-scale study in the orkut social network', *KDD'05*, pp.678–684.

Srivastava, J., Cooley, R., Deshpande, M. and Tan, P. (2000) 'Web usage mining: discovery and applications of usage patterns from web data', *SIGKDD Exploration Newsletter*, Vol. 1, No. 2, pp.12–23.

Sun, J., Faloutsos, C., Papadimitrious, S. and Yu, P.S. (2007) 'Graphscope: parameter-free mining of large time-evolving graphs', *KDD'07*, pp.687–696.

Symeonidis, P., Nanopoulos, A. and Manolopoulos, Y. (2010) 'A unified framework for providing recommendations in social tagging systems based on ternary semantic analysis', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, No. 2, pp.179–192.

Tang, J., Sun, J., Wang, C. and Yang, Z. (2009) 'Social influence analysis in large-scale networks', *KDD'09*, pp.807–816.

Tang, J., Wu, S., Sun, J. and Su, H. (2012a) 'Cross-domain collaboration recommendation', *KDD'12*, pp.1285–1293.

Tang, L., Liu, H. and Zhang, J. (2012b) 'Identifying evolving groups in dynamic multimode networks', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No. 1, pp.72–85.

Tang, L., Wang, X. and Liu, H. (2012c) 'Community detection via heterogeneous interaction analysis', *Data Mining Knowledge Discovery*, Vol. 25, No. 1, pp.1–33.

Tang, L. and Liu, H. (2010) *Community Detection and Mining in SocialMedia*, Morgan & Claypool.

Tantipathananandh, C., Berger-Wolf, T. and Kempe, D. (2007) 'A framework for community identification in dynamic social networks', *KDD'07*, pp.717–726.

Tso-Sutter, K., Marinho, L. and Schmidt-thieme, L. (2008) 'Tag-aware recommender systems by fusion of collaborative filtering algorithms', *SAC'08*, pp.1995–1999.

Wang, C., Satuluri, V. and Parthasarathy, S. (2007) 'Local probabilistic models for link prediction', *ICDM'07*, pp.322–331.

Wu, J., Xiong, H. and Chen, J. (2009) 'Adapting the right measures for k-means clustering', *KDD'09*, pp.877–886.

Xu, G., Gu, Y., Dolog, P., Zhang, Y. and Kitsuregawa, M. (2011) 'Semrec: a semantic enhancement framework for tag based recommendation', *AAAI'11*, pp.1267–1272.

Xu, T., Zhang, Z., Yu, P.S. and Long, B. (2008a) 'Dirichlet process based evolutionary clustering', *ICDM'08*, pp.648–657.

Xu, T., Zhang, Z., Yu, P.S. and Long, B. (2008b) 'Evolutionary clustering by hierarchical Dirichlet process with hidden Markov state', *ICDM'08*, pp.658–667.

Yang, C., Harkreader, R., Zhang, J., Shin, S. and Gu, G. (2012a) 'Analyzing spammers social networks for fun and profit', *WWW'12*, pp.71–80.

Yang, X., Steck, H. and Liu, Y. (2012b) 'Circle-based recommendation in online social networks', *KDD'12*, pp.1267–1275.

Yang, T., Chi, Y., Zhu, S., Gao, Y. and Jin, R. (2009) 'A Bayesian approach toward finding communities and their evolutions in dynamic social networks', *SDM'09*, pp.990–1001.

Yu, L., Pan, R. and Li, Z. (2011) 'Adaptive social similarities for recommender systems', *RecSys'11*, pp.257–260.

Zhang, Y., Yu, J. and Hou, J. (2006) *Web Communities: Analysis and Construction*, pp.257–260, Springer, Berlin, Heidelberg.

Zhang, Z., Zhou, T. and Zhang, Y. (2010) 'Personalized recommendation via integrated diffusion on user-item-tag tripartite graphs', *Physica A: Statistical Mechanics and its Applications*, Vol. 389, No. 1, pp.179–186.

Zhao, Y., Levina, E. and Zhu, J. (2011) 'Community extraction for social networks', *Proceedings of the National Academy of Sciences of the USA (PNAS)*, Vol. 108, No. 18, pp.7371–7326.

Zheng, Y. and Zhou, X. (2011) *Computing with Spatial Trajectories*, Springer, Berlin, Germany.

Zheng, Y., Zhang, L., Ma, Z., Xie, X. and Ma, W. (2011) 'Recommending friends and locations based on individual location history', *ACM Transactions on the Web*, Vol. 5, No. 1, Paper 5.

Notes

1 http://www-personal.umich.edu/~mejn/netdata/.
2 http://snap.standford.edu/data/.
3 http://www.delicious.com.
4 http://www.medworm.com.
5 http://www.citeulike.org.