

# A Survey of User Modelling in Social Media Websites

Ahmad Abdel-Hafez<sup>1</sup> & Yue Xu<sup>1</sup>

<sup>1</sup> School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, Australia

Correspondence: Ahmad Abdel-Hafez, School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, Australia. E-mail: a.abdelhafez@qut.edu.au

Received: June 29, 2013 Accepted: August 1, 2013 Online Published: September 11, 2013

doi:10.5539/cis.v6n4p59

URL: <http://dx.doi.org/10.5539/cis.v6n4p59>

## Abstract

With the widespread of social media websites in the internet, and the huge number of users participating and generating infinite number of contents in these websites, the need for personalisation increases dramatically to become a necessity. One of the major issues in personalisation is building users' profiles, which depend on many elements; such as the used data, the application domain they aim to serve, the representation method and the construction methodology. Recently, this area of research has been a focus for many researchers, and hence, the proposed methods are increasing very quickly. This survey aims to discuss the available user modelling techniques for social media websites, and to highlight the weakness and strength of these methods and to provide a vision for future work in user modelling in social media websites.

**Keywords:** user modelling, social media, semantic enrichment, and dynamic user modelling

## 1. Introduction

With the widespread of social media websites in the internet, and the huge number of users participating and generating infinite number of contents in these websites, the need for personalisation increased and became a necessity. One of the major issues in personalisation is building users' profiles; this challenging process has been attracting researchers' attention in the last decade. Researchers aim to provide solid users models that can deliver accurate users' preferences, which can be used by applications in order to enhance usage experiences in the widespread social media websites. Building users' profiles depends on many elements; such as the available data, the application domain they aim to serve, the representation method and the construction methodology etc. To this end, current researches provided different directions and methods in building users' profiles.

Social media websites currently represent the soul of the internet for millions of people, and they are spreading more and more. And because social media websites are diverse and have several types of data, the user profiling methods were also diverse and sometimes domain dependant. For example, profiling users in social network websites are different than product rating websites or social bookmarking websites, and that's because of the existence of different elements and data about users in these websites. Even within one category such as social networks, modelling users will have different methods between different websites, i.e. Twitter depends on micro-blogs, while Facebook has many other elements such as sharing contents, joining groups and pages, besides to the commenting system and status update.

In this work, we aim to discuss the current researches conducted in the area of user modelling for social media and provide an overview of future research directions. As we mentioned above, many researches were conducted lately, which have focused on modelling users for social media websites, and for this reason we believe that these works require grouping and analysis in order to figure out where the research is heading to in this area and what is the possible future development. The main contribution of this paper is to highlight the available user modelling techniques for social media websites, to highlight the weakness and strength points of these methods, and to provide a vision for future work in user modelling.

## 2. Social Media

In the last decade, there was an enormous amount of data published on the internet on a daily basis by all kind of users all around the world. Social media websites were the pillar of this evolution since they provided web users with the frameworks required to establish collaborative works and generate web contents. In this section, we will

introduce the definition and elements of social media, and then we will discuss the problems and issues arise with it.

### 2.1 Social Media Definitions

In the free on-line dictionary, social media has been defined as “web sites and other online means of communication that are used by large groups of people to share information and to develop social and professional contacts”. On the other hand, Ahlqvist et al. (2008) provided a definition that is built on three key elements: content, communities and Web 2.0. Their definition was; “social media refers to the interaction of people and also to creating, sharing, exchanging and commenting contents in virtual communities and networks”. More extended definitions were introduced later, for example, Kietzmann et al. (2011) included mobile applications besides web-based applications. Social mentioning that “social media employ mobile and web-based technologies to create highly interactive platforms via which individuals and communities share, co-create, discuss, and modify user-generated content”. While Kaplan and Haenlein (2010) defined social media as “a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of User Generated Content”. In order to emphasize the web part of social media in this paper, we define social media website as: “a web site that provides an interactive platform, which facilitates communication between people or creating and sharing User Generated Contents (UGC), including collaborative works, social networks, blogs, contents sharing, social bookmarking, virtual worlds and rating websites”. In our work we will use the term social media to indicate social media websites.

Nowadays, social media is used by millions of people, and still growing exponentially, which consequently has exponential impact on the amount of UGC, and on the social connections between people. Basically, the existence of social media encouraged people to give their opinions more freely and participate more in many aspects of life, such as politics, and its impact is very clear in what is called the “Arab Spring” (Saleh, 2012). Besides its huge impact on businesses, online users' opinions can enforce the success of a product or destroy the reputation of another. On the other hand, social media has many advantages to users, starting from self-entertaining and meeting new people, to the availability of a huge source of information on hand about almost anything they might think of.

As we have mentioned in the social media definition above the social media website should facilitate communication between people and allow them to create contents. To this end, many websites nowadays, which are not categorized as social media in their nature, are embedding sociality services into their original website activities. For example, news websites now allow users to comment on news articles and rate or share them, which make them under the scope of social media websites; while the core of their website still not changed (Ahlqvist et al., 2008). A common part in all social media websites is the ability to collect data from users and use it in order to build users' profiles, which may contain their social behaviour, and their general interest. This information is very useful in personalization in general.

### 2.2 Social Media Classification

Kaplan and Haenlein (2010) classified social media websites by social presence/media richness and self-presentation/self-disclosure; we will use their classification in this paper.

Table 1. Classification of social media presented by Kaplan and Haenlein (2010)

		Social presence/Media richness		
		Low	Medium	High
self-presentation /self-disclosure	High	Blogs and Rating and Reviews websites and social bookmarking	Social Networks	Virtual worlds
	Low	Collaborative work	Content sharing	Virtual game worlds

Table 1 shows the different classes of social media websites.

- **Social Networking:** They allow users to establish connections or relationship with other users in the network, like friendship in Facebook, or follow relation in twitter (Zhou, Xu, Li, Jøsang, & Cox, 2012). In the former one both users must provide acceptance for the relationship to be created, while in the second one you don't need a user permission in order to follow him. Some of these websites are not

general, but rather they impose one kind of social connections, like professional connections in LinkedIn (Kietzmann et al., 2011). Social networks nowadays are multiuse websites, where you can communicate with friends, read news, play games, join interest groups, share media files, and much more.

- Collaborative Works: They allow users to participate and communicate with each other in order to build huge useful databases such as Wikipedia, where users all around the world work together in order to achieve an ultimate goal, which is the build of free huge encyclopaedia ever made by humans.
- Content Sharing: They provide users with the suitable platform to share contents such as sharing videos in YouTube, or sharing photos in Flickr. They allow users to rate or comment on contents, and also to attach tags to these contents.
- Blogs: They provide a more open environment for users' text comments and discussions about any topic they are interested in. They are very popular on the internet as they are easy to maintain and manage. Stackoverflow website represents a modern example of forums, where it provides expertise exchange in computer programming in the form of questions and answers.
- Ratings and Reviews: They offer users with a sole chance to share their opinions about products with other users using ratings and textual comments, like e-Bay, Amazon, C-Net, and Epinions. Consequently, they provide a great potential for both users and companies to learn more about products and their actual pros and cons after they were used by customers. Nowadays, they proved to have a huge impact on customers' decision making process.
- Social Bookmarking: They provide the opportunity for users to add, annotate, edit, and share bookmarks of web documents (Noll & Meinel, 2007). Besides users can vote on websites and rank them according to users' preferences; such as Delicious, and Reddit. They only provide a reference to the bookmarked website, unlike content sharing which provide the resources themselves.
- Virtual Worlds and Virtual Game Worlds: They provide a simulated environment where users' can interact to each other to form online communities; usually they are represented in 3D graphics, such as Second Life, and IMVU (This class is out of the scope of this paper).

### 2.3 Issues and Problems

With the massive amount of available UGC on the Web, and the wide range of services provided by social media websites, many issues arise associated with businesses and their communication with their potential customers, and also with users' lifestyle and interaction with social media elements (Ahlqvist et al., 2008). On the other hand, users are concerned with issues such as privacy, identity theft, addiction, and spread of bad information. While these issues are very important in social media, but they are out of our scope in this survey, we are more concerned with issues related to the usage experience of social media websites.

One of the usage issues of social media websites is that social media websites become more difficult to access proportional to the size of available data. For example, when the number of your friends, liked pages, and Apps increase intensely in Facebook, it becomes more difficult to follow all the news feeds from them, so there is a higher chance that you will miss interesting news feed from a friend due to the huge number of unrelated feeds from other elements. Another example is in YouTube, where it becomes more difficult to find interesting video in between the billions of the available ones. This problem is believed to be solved by personalisation; as every user will see more items that he is interested in, which in turn requires unique users' profiles to keep users' preferences (Ahlqvist et al., 2008).

Recently, another issue has been addressed by a couple of recent published work (Ahmed, Low, Aly, Josifovski, & Smola, 2011; Gueye, Abdessalem, & Naacke, 2012; Li, Yang, Wang, & Kitsuregawa, 2007; Xiang et al., 2010) that is referred to as the dynamicity problem. We define it as the effect of time on user's preferences and how it can be reflected in their profiles. Modelling dynamic users' profiles can help in providing more quality services for users, such as, providing the right ad at the right time by emphasising the short term users' interests (Ahmed et al., 2011). Moreover, recommender systems may use the dynamicity feature in order to enhance the predictions accuracy of users' ratings and in turn enhance the quality of recommender systems (Gueye et al., 2012).

In regards to businesses, companies nowadays understand the value of social media websites and they are trying to make advantage of them due to their importance on the progression of the business in future. The problem they face is how to achieve this goal effectively, especially with the diversity of nature of the available social media websites; one research suggested to treat them as an ecosystem of related elements when you develop a

social media strategy rather than treating them as standalone systems (Hanna, Rohm, & Crittenden, 2011). Other work focused on more detailed issues; such as how to treat negative spreading opinion about your product on social media (Noble, Noble, & Adjei, 2012), or how to make use of customers stories to enhance your product or service (Gorry & Westbrook, 2011). In general, the existence of social media affected the communication process between the company and potential customers, and how users' opinions will participate in the innovation processes and new products and services development (Ahlqvist et al., 2008). Companies can benefit from using distinct users' profiles in the way they communicate with potential customers; for example, if a specific company is producing a diverse range of products, it is more convincing to offer the appropriate product for each customer according to their needs and interests (Zhou et al., 2012). On the other hand, users' profiles can provide an organized method to make use of social media diverse data in determining the directions of Research and Development (R & D) in order to fulfil customers' requirements.

As a summary, social media has a wide range of issues and problems that affect both business and Web users. A common solution for some of these issues lies in establishing of well-constructed users' profiles, as they can provide a general perspective of users' interests, and a base for services personalisation, whether from a website or company side.

### 3. User Profiling in Social Media

As defined by Zhou et al. (2012) "User profiling is the process of acquiring, extracting and representing the features of users". The profile can be used to present more relative content to each user and they usually contain users' basic information; such as age, gender, country ... etc., and keywords or concepts that represent users' interest. More sophisticated profiles may contain users' behaviour information; such as sequence of clicks and time spent on pages, this can be useful in personalization as well. Recently, some researchers suggested using users' social information in building users' profiles; such as social connections with other users or groups and pages, and also social behaviours like shares, clicks, and likes between users (Kim, Ha, Lee, Jo, & El-Saddik, 2011; Lu, Lam, & Zhang, 2012; Tao, Abel, Gao, & Houben, 2012). Social information is believed to be useful in enhancing many predictive results of different applications (Ma, Zhou, Liu, Lyu, & King, 2011; Mezghani, Zayani, Amous, & Gargouri, 2012; Yang, Steck, & Liu, 2012; Yu, Pan, & Li, 2011). Many efforts have been made previously that provided well organized and detailed surveys about personalisation in the web, and user profiling (Anand & Mobasher, 2005; Gao, Liu, & Wu, 2010; Gauch, Speretta, Chandramouli, & Micarelli, 2007). In this survey, we will focus on the latest trends in user modelling research related to social media and we will provide a view for future research in this area.

Figure 1 shows the steps of building users' profiles, in general. It is an extension to the figure introduce by Gauch et al. (2007) which does not contain the enrichment process. The first step is the data collection, which gathers users' data from social media websites including filled in forms data, log file data, and connections with other people in the system. The second step is the profile construction, where the users' interests will be extracted and represented using different methods; weights also will be embedded with every interest showing the degree of interest. The result of this step will be a user profile represented as a vector, graph, or hierarchy. The graph and hierarchy based profiles require an additional step in the methodology in order to extract relationships between keywords. The Enrichment process aims to add more related keywords to the profile in order to enhance the final prediction results; many sources can be used in order to extract the extra keywords; such as WordNet synsets (synonyms sets), Web sites like Wikipedia or news articles, and like-minded users or friends profiles. Finally, the profile is ready to be used by different personalisation based applications; such as recommender systems, ads generations, e-commerce, etc.

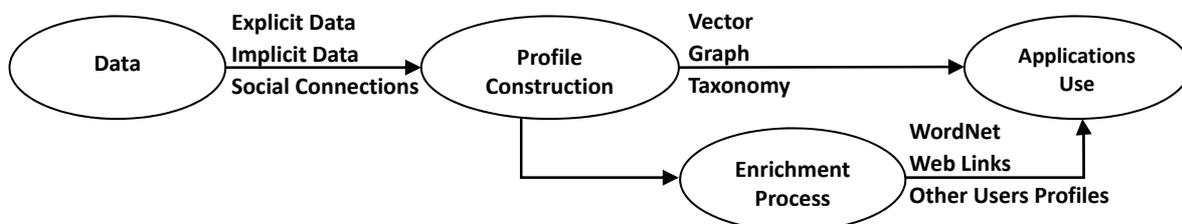


Figure 1. User profile construction process

### 3.1 Data Sources

Basically, the collected data depends on the nature of the Website used and the target application. In general, we can obtain explicit, implicit, and social data.

#### 3.1.1 Explicit Data

Explicit data is given directly by the user; such as demographic information, comments, search queries, and ratings (Mezghani et al., 2012). Some researchers use users' comments and posts directly to extract keywords to represent users' interests (Hannon et al., 2010; Lu et al., 2012), while others directly use the rated items as indication of users' interest (Ma et al., 2011). On the other hand, the use of demographic information similarities can generate interests for new users who still have no rating history in order to solve the cold start problem, which is when a user visit a website for the first time (Kim et al., 2011). Tags are also commonly used as direct interest keywords when they are attached by the user to a web content, or using social bookmarking websites (De Pessemier, Deryckere, & Martens, 2009; Hannon et al., 2012; Hung et al., 2008; Michlmayr & Cayzer, 2007).

#### 3.1.2 Implicit Data

In contrast, the implicit data refers to the inferred data from users' behaviour and they could be acquired by studying user clicks, transactions, and navigation data; for example, when a user clicks on a link and open a web page we can extract the page title as the user's interest, or we can extract keywords from the page content if the user has spent time larger than a pre-defined threshold on this page (Das, Datar, Garg, & Rajaram, 2007). Some researchers considered user clicks as explicit data, as it is intended by the user, while implicit data is the data that does not involve user interaction with the computer, such as linger time, which is the time spent on a specific Webpage, which can be extracted from the user log data, or mouse over and eye movement (Riggs & Wilensky, 2001).

#### 3.1.3 Social Connections Data

Social data represents relationships or interactions among users. The relationships can be bidirectional which requires the acceptance of both connected users or unidirectional such as the follow/followed connections in Twitter. Social network data can be represented as a graph, and the graph analysis can help in identifying user communities in the network. In general, social graphs are used in many researches (Bhuiyan, Xu, Jøsang, Liang, & Cox, 2010; Ma et al., 2011; Yang et al., 2012) as a trusted community for the user, which can be treated as like-minded group of users. This method may replace or work side by side with the nearest neighbour method which depends on similarities between users in order to identify like-minded users. They are also used to enrich users' profiles with more interests' words assuming that a user will be interested in a common topic that interest his friends, or his similar taste friends (Hannon et al., 2012).

### 3.2 Keyword-Based User Profile Representation

A keyword-based user profile is usually represented as a vector, which is a simple and common representation used to represent user profile as pairs of concepts and related weights. The concepts represent users' interests and the weights represent the degree of interest. Values can be binary (0 or 1), to indicate behaviours such as; purchase or not, clicked or not, or they can be integers; such as items' ratings or term frequency (*TF*) (Barla, 2011). They can also be real numbers that represent weights; which can be calculated using several methods such as term frequency multiplied by inverse document frequency (*TF*×*IDF*) Equation (1).

$$TF \times IDF(c) = TF(c) \times \log \frac{n_i}{N} \quad (1)$$

*TF*: is the frequency of the concept,  $n_i$  is the number of documents that contains this concept, and  $N$  is the total number of documents.

Suppose we have  $N$  users, each is identified by number of concepts  $C_k$ ; then the user's profile is represented as a vector  $P(U_i) = \langle W_{i1}, W_{i2}, \dots, W_{im} \rangle$  Where "m" represents the dimensionality of the vector and  $W_{ij}$  is the weight for the j-th concept (Salton, Wong, & Yang, 1975). As an abstract, users' profiles are represented as  $P(u_i) = \langle (c_j, w(u_i, c_j)) | c_j \in C \rangle$ ,  $u_i \in U$  Where  $C$  and  $U$  are set of concepts and users respectively, and  $w$  is the weighting function.

Other representations for users' profiles include graph-based and hierarchy-based profiles. These two kinds of profiles consist of nodes and arcs. The nodes usually represent the keywords or topic of interest and the arcs represent the relationships between these nodes. In some cases it was proposed that these arcs must be associated with weights, which are used to define the strength of the relation between any two nodes.

### 3.3 User Profile Construction

In this section we will review the methods used to construct users' profiles which includes methods for extracting user's interest keywords and their associated weights, and then extracting relationships between keywords in case of graph and taxonomy based representations:

#### 3.3.1 Traditional Bag of Words (BOW)

BOW is a simple method to generate the keywords which will represent user's interests in the profile. Usually, this method is used with systems that depend on explicit data like micro-blog text. The BOW are collection of the words used in user's text and their frequency are the weight, or the more complex  $TF \times IDF$  method is used to calculate the weight of each word. Hannon et al. (2010) used this method to represent Twitter users' profiles. Similarly, Chen et al. (2010) did the same to build the profile using  $TF \times IDF$  weighting, he also built followee profile by collecting words from followees tweets and choosing the highest 20% TF scores and omitting words that appear in one followee profile only. And they call the resulting set of words high-interest words. On the other hand they model URLs by the words used to describe them on users' tweets and then they use cosine similarity to decide whether this URL is in the scope of user's interest or not. However, The main problems with this method is Polysemy, which is the presence of multiple meanings for one word, and Synonymy, which indicate that relevant information can be missed unless the exact keyword exist in the profile (Lops et al., 2009). Besides, other methods proved to generate better quality users' profiles, and that explains why it is rarely used nowadays by researchers.

#### 3.3.2 Concepts Based

This method is very common in user profiling, where concepts are extracted from users' data in several ways. Kim et al. (2011) used a text mining method, which involves three stages; extracting terms, mining frequent patterns, and pruning patterns. They used implicit sources of data, such as clicked, viewed, and bookmarked items, and extracted terms from these contents. Authors used the  $TF \times IDF$  method for terms weighting and then they mine for frequent term patterns. In the last step they removed patterns containing unnecessary terms from the set of frequent term patterns. Lu et al. (2012) used Wikipedia as a rich external source of data in order to extract concepts from users' tweets on Twitter. First they represented each concept of Wikipedia as a vector of pairs of words and weights using  $TF \times IDF$  method, and likewise they do with each tweet, then they extract relevant concepts using Explicit Semantic Analysis (ESA) by computing semantic relatedness between Wikipedia concept vector and tweet vector. Authors also add a vector of social connections, which includes other users and affinity scores calculated by counting number of tweets that reply, re-tweet, or mention between two users. Semeraro, Degemmis, Lops and Basile (2007) targeted solving the problems of Polysemy and Synonymy by using a synset-based vector space representation, called bag-of-synsets (BOS). They apply (Word Sense Disambiguation) WSD procedure to documents and extract synset for each word using the context words, defined as a set of words that precede and follow a given word, then the user profile is built as a synset vector, rather than a word vector, and the weight vector represent the frequencies of the synset. Other weighting systems can be used too. Authors used a Naïve Bayes text categorization algorithm to build profiles as binary classifiers (user-likes vs. user-dislikes).

#### 3.3.3 Tag Based

Many social media web sites provide social tagging capacity to users; they enable them to annotate items with tags of their choice; such as Flickr or Delicious. These annotation processes are represented as quadruple representation of user-tag-resource-relation; which is called Folksonomy (folks taxonomy) (Wall, 2007). The relation part might indicate the time when a tag assignment was created  $F = \langle U, T, R, Y \rangle, Y \subseteq U \times T \times R$ .

Hung et al. (2008) used tags provided in Flickr and Delicious to build user's profile, They introduced two means in the users' profile; the personal view and the social view. In the former part they only consider the tags assigned by the user himself, while on the second part, they consider tags assigned by user's social contacts. On their proposed weighting system, they assume that the first tag of a specific bookmark is more relevant than the second tag and should get more weight, and so on. Hence, they apply exponential decreasing function up to the tenth tag, assuming that the rest of the tags will have similar weights. On the other hand, Abel et al. (2011c) introduced Mypes; which is a cross-system user modelling depending on collecting tags from different social tagging systems and mapping them together using simple rules to convert service specific vocabulary to common vocabulary. The major challenge they have faced was connecting different user's accounts on different websites to each other. In order to solve this problem they used Google social graph which provide this service for users who linked their accounts via their Google profile. Hannon et al. (2012) used (<http://listorious.com>), a category database which maintain Twitter curated lists, hand-annotated with topical tags by users, in order to extract tags

about each user, which is a set of all tags that represent all the lists the user belong to. Lops et al. (2009) presented a method similar to the one presented by Semeraro et al. (2007), except that they used tags instead of words. They also includes social tags besides to personal tags in users' personal profile, and by social tags they mean the set of tags provided by all the users who rated a specific item, that is rated by the user and the personal tags are the tags provided by the user himself. They create two sets of synsets obtained by disambiguating the personal tags set, and the social tags set, calling them Semantic Personal Tags, and Semantic Social Tags.

#### 3.3.4 Topics Based

Topic modelling techniques are used in order to represent user interest as topics rather than keywords; this method is argued to provide a better performance (Ahmed et al., 2011; Weng, Lim, Jiang, & He, 2010; Zhong, Fan, Wang, Xiao, & Li, 2012). Ahmed et al. (2011) model users' interests as latent topics based on latent Dirichlet Allocation (LDA), where they maintain two distributions, users' distributions over topics and topics' distributions over words. They used user queries to collect the words of interest for the user in order to enhance advertising targeting. In their proposed model TVUM (Time-Varying User Model) they divided user actions into epochs, where actions (represented by words) inside each epoch are modelled using fixed-dimensional hierarchical Polya-Urn model of LDA. This model indicates that previously expressed interests are more likely to be expressed by the same user or other users. Their aim was to filter out external effects from users' profile, assuming that they are not part of users' interests. Zhong et al. (2012) presented (ComSoc) model to transfer user's behaviour over composite social networks. They introduce a term of users' distribution over networks, which represent the probability of how much a user is influenced by a given network. At first, a network is drawn for each user from a Dirichlet distribution, then, for every interaction of a given user, a social network is drawn from a Multinomial distribution. Each user can adopt relationship from different sub-networks individually according to their similarities to others. On the other hand, Weng et al. (2010) introduced TwitterRank, a system that relies on the topics of tweets; their goal was to identify topic based influential micro-bloggers. They were motivated by the fact that twitterers have different level of experience in different topics; consequently, they will have different influence in each topic.

#### 3.4 Semantic Enrichment

The semantic enrichment process aims to enhance the scope of the words used to represent users' interests, and to provide a prediction for new interests of the user that were not explicitly mentioned by him. Basically, the dependence on users Micro-Blogs; such as tweets on Twitter, in order to build users' profiles provide a narrow but important source of information as the text is very short (limited to 140 characters in Twitter) (Abel, Gao, Houben, & Tao, 2011a). Moreover, users' modelling methods that use enrich profiles in order to avoid the cold start problem by providing a more detailed picture of user needs. On the other hand, enrichment can enhance CF recommendations by providing more accurate similarity results between users; for example, two users' profiles could not be recognized as similar, but after enrichment they appear to be similar (De Meo, Quattrone, & Ursino, 2010). Researchers provided ideas on how to enrich the semantics of micro-blogs in order to build users' models. They suggested mapping micro-blogs posts into many other sources and use words from these sources in order to semantically enrich users' profiles with new related words; such as using:

##### 3.4.1 WordNet

WordNet provides synsets for words, Degemmis et al. (2007) used WSD algorithm using context to discover the correct meaning for each word and then enrich the profile with the synset of the word (Degemmis, Lops, & Semeraro, 2007), these words can represent the same meaning and may be used interchangeably. Similarly, Lops et al. (2009), Semeraro et al. (2007) include the WordNet synsets in users' profile after performing a WSD process, they link a synset for each word and calculate frequencies of synset occurrences. Abel et al. (2011c) enrich user profile with metadata that denote the top-level categories for words extracted from WordNet. If the words are not contained in WordNet they use DBpedia. The purpose of attaching a category for each word is to provide the word sense, and it is used later on in their system to filter tags according to the desired category.

##### 3.4.2 Wikipedia

Lu et al. (2012) extracted concepts from user tweets and expand them by finding related concepts using Markov random walk on the Wikipedia graph assuming that strongly related concepts will be in the scope of users interest even it doesn't appear in his profile. For example, if a user is a fan of Apple products, he will be interested in new products from Apple even if it doesn't appear in his profile. Similarly, Xu and Orad (2011) used Wikipedia as an external source to enrich micro-blog posts for the purpose of topical clustering. Wikipedia has been also used through DBpedia; which is "a crowd-sourced community effort to extract structured information from Wikipedia and to make this information available on the Web". Jadhav et al. (2010) suggested

identifying concepts that are related to a specific topic in order to form the semantic keyword clusters using DBpedia, then these clusters can be used to enrich users' profiles that contain words of these clusters.

### 3.4.3 Web Links

Unlike using WordNet, enriching users' profiles using news articles requires an extra step in order to function, which is linking the user's text message to the right news article. This process can be done using URL-based strategies or content based strategies. In the former one, the user explicitly adds the URL of the news article in his message, or it may appear in another message a user replied to or forward. While the later one is more difficult because it requires measuring similarity value between user text and news article, which can be represented using bag of words, or hash-tags, or entity based methods. In their work, Abel et al. (2011b) used the  $TF \times IDF$  method to measure similarity between user tweet and news articles and chooses the most similar ones to build the link. Santos and Nguyen (2009) created interest set based on the intersection of retrieved relevant documents, rather than using the user's search query as interest words. Jadhav et al. (2010) used Google Insights for Search, which provides the trends of searches in specific location and time to build the semantic clusters in the cases where words are not found in DBpedia.

### 3.4.4 Socially Connected Users

Using other users' data has been proven to be useful in enriching users' profiles with users' potential interests. Chen et al. (2010) used followees tweets in order to discover topics of interests for a user, they collected all the word in followees' tweets then find frequency of each word, and they use only the top 20% of these words, where the weight is the number of users who mentioned this word. Instead, Hannon et al. (2010) used followers and followees tweets in order to expand user's profiles, and they weighted the extracted words using TFIDF method. They proved the effectiveness of their method by evaluating followees recommendations to the user in a real time Web. On another work authors used tags instead of tweets, extracted from Listorious, database of lists created by twitter users and annotated with a set of topical tags, they described a multi faceted user model. Their model partition the user tag-space into seven disjoint regions showing all alternatives of tags' overlapping between user and his followers and followees (Hannon et al., 2012). Hung et al. (2008) represented a similar idea of enriching user's profile by adding tags from her friends' profiles, except that they used only the profiles that share one or more tags or resources with the main user profile.

### 3.4.5 Like-Minded Users

Kim et al. (2011) proposed a method to enrich a user's profile from his like-minded users' profiles. As a first step for their method, they attempted to discover the k-nearest neighbours of a user. They used the cosine similarity method to measure the a similarity value between a user and every other users, this method quantifies the similarity of a pair of vectors according to their angle generating a value between 0 and 1; where the higher the resulted value the more similar the vectors are. After determining the nearest neighbours for a user, their profiles will be used to enrich his profile, assuming that he will have similar interests as them. The basic idea is that the pattern found in more users' profiles, contribute more in enrichment process. Authors indicated that this enrichment process is particularly effective to solve the cold start problem, where user's profile is short of enough interest terms and patterns.

## 4. Dynamic User Models

User's profile dynamicity refers to the change that happen to the user's interests over time. Researchers provided many methods in order to reflect the changed interests over time in order to build more accurate profiles that can be more useful when used with applications. Basically, the idea in most of the proposed works was to add volatility factor to weighting methods, which will reduce the interests' weights by time if they are not used by the user until they disappear. Santos and Nguyen (2009) incorporated a fading function to make the irrelevant interests disappear by time. In contrast, Michlmayr and Cayzer (2007) introduced the adaptive user profiles by adding Evaporation and reinforcement elements. Evaporation is to reduce the tags weights, by removing a small percentage, each time the profile is updated, and Reinforcement is to increase the weight of edges that appear again while it is already existed in the graph. They adopt an iteration-based graph visualisation algorithm which allow them to identify active and not active interests, as well as into long-term, mid-term, and short-term ones.

Ahmed et al. (2011) provided dynamicity at three levels in their model, the global distribution over interests, the user-specific distribution over interests and the topic distribution over words. They also find short term and long term interests, and combine them using weighted average to get the expected user-specific's popularity over interests at a specific time. Yu et al. (2012) explained two cases to update concepts weight "life-time", the first if the concept that appear in the new session is currently available in the model, and in this case the new weight

will be the average of both weights, which will increase or at least keep the weight of the concept. The other case is if the concept didn't appear in the session and is currently available in the model. In this case the weight for the concept will be reduced, the percentage of reduction is calculated based on two factors; the semantic difference degree between existing UP-CR file and the new session semantics file calculated using cosine method, and temporal difference degree which is computed based on the number of sessions between concepts last arising and the updating moment. In this case, the updated life time of a node decreases, and if this node does not arise in the next several sessions, its life time will decrease greatly until it is deleted from UP-CR.

## 5. Future Directions

There are many work efforts that is required to be addressed in regards to user profiling in social media websites. Until now, researchers have been focusing on traditional profiling strategies without considering the diversity of elements provided by different social media website. Besides, profiles enrichment has not been given an adequate attention. Moreover, the need for dynamic and more intelligent profiles requires more attention in order to achieve the best results from users' profiles. In this part we will focus on the next step in the future of user profiling and what is needed to be done.

### 5.1 More Dynamicity

Current researchers have focused on the dynamicity feature of users' interests, assuming that dynamic users' profiles will produce more accurate results at the application level, such as recommender systems. Dynamicity mainly was implemented by adding the time element into the equations of calculating weights, which represents users' degree of interest. In this section we will suggest future changes to enhance the dynamicity influence on accuracy of applications. First of all, fading or evaporating of users' interests in the profiles has been implemented many times in the current researches, on the other hand, using different fading variable for every topic or cluster of interests may be useful and may reflect more accuracy in the users' profiles. The assumption behind this idea is that users' interests on different topics do not evaporate or fade at the same speed. This issue also can be solved by defining a long term and short term interests as suggested in (Ahmed et al., 2011) and then combine them together in order to generate a dynamic weights. However, authors didn't explain in details how they determine long-term interests and whether they change by time or not. We believe that more work efforts can be done towards this issue.

Depending on the theory of revisit (Tauscher & Greenberg, 1997), we can assume that users usually have groups of similarly browsing behaviours in different sessions. We can call this the users' mood, where a user can be interested in a specific part of his overall profiles' interests in different sessions. Depending on the previous user's behaviour (visited pages, sequences of actions in a single website, etc.); we can cluster sessions with similar behaviour together, and model different weighting systems for each one of them in the users' profile. Different moods can be detected per user, where the interest weights vary in each mood. In the beginning of a new session a dynamic weighting will be provided depending on users observed behaviour. Finally, the absence of user action against specific elements in webpage, and the sequence of click streams in some websites have not been well studied in research. We believe that giving more attention to these parts of implicit data can provide more dynamic profiles.

### 5.2 More Enrichment

Enrichment process was added in order to enhance users' profiles by making them rich with related words that were not mentioned originally by the user himself. Most of the research papers who used this step, they choose only one source for enrichment process, such as web links, Wikipedia, WordNet, friends profiles or similar users' profiles. However, there was no study that evaluates each one of these semantic enrichment methods of users' profiles and provides a comparison between them, and highlights the strength and weakness of each one of them in several application domains. Moreover, trying to combine different sources may even enhance users' profiles even more. On the other hand, discovering other sources for enrichment is a possibility, such as using search engines returned results, or the content of multimedia files such as converting speech to text and using the text in the enrichment process. At the end, can we determine the best combination of enrichment sources that provides the best results in any application domain?

### 5.3 More Comprehensive

In Table 2, we tried to summarize some of the latest work showing all the aspects of users modelling. The differences between the suggested methods can be clearly noticed from the table. To this end, we can conclude from the table that most of the proposed works have ignored one or more from the important aspects of users' modelling. For example, Kim et al. (2011) didn't embed dynamicity into his modelling method, where Ahmed et

al. (2011) ignored social relations which can affect his results. On the other hand, both of them ignored the relationship between interest topics or keywords, and whether this relationship is important in providing more accurate users' profiles or not. We believe that more effort must be given in order to build more comprehensive models that can use all kind of data sources and provide enriched dynamic profiles, without ignoring relationships between interests.

Table 2. Summary of proposed user profiling methods in social media

Representation Method	Types of Data used			Keywords Extraction Method	Weighting Method	Enrichment Source	Enrichment Method	Profile Dynamicity	Author
	E	I	S						
Vector	✓	✓		BOW from (Tweets)	TF-IDF	Followers, and followees tweets	TF-IDF	-	(Hannon et al., 2010)
						Followee tweets	TF		(Chen et al., 2010)
	✓	✓	✓	Frequent Pattern Mining	TF-IDF	Personalized term patterns of like-minded users	cosine similarity	-	(Kim et al., 2011)
	✓		✓	Matching Tweets to Wikipedia Concepts	Explicit Semantic Analysis (ESA)	Wikipedia related concepts	Markov Random Walk	-	(Lu et al., 2012)
	✓			Naïve Bayes text categorisation	TF-IDF	WordNet Synset for each word	Word Sense Disambiguation	-	(Semeraro et al., 2007)
					TF	Followers and followees tags	TF	-	(Hannon et al., 2012)
	✓	✓		Tags	Average of exponentially decreased weights reliant on tag sequence per user)	Friends tags	Average of weights calculated using exponential decreasing function depending on tag sequence per user	-	(Hung et al., 2008)
	✓	✓		Topic Modelling	Polya-Urn Representation of latent Dirichlet Allocation (LDA)	-	-	Interest topics in different epochs average weight	(Ahmed et al., 2011)
	✓			Items clusters (Similarly rated items)	Average ratings per cluster	-	-	Dynamic item clustering	(Wen & Zhou, 2012)
	✓			Any item clustering method	Average ratings minus bias value per cluster	-	-	-	(Gueye et al., 2012)
Representation Method	E	I	S	Keywords Extraction Method	Weighting Method	Relations Between Nodes Extraction	Relation Weights	Profile Dynamicity	Author
Graph	✓	✓		Frequent Pattern Mining	$TF / n * \Sigma$ Browsing positions ratios	- PMI	PMI * $\Sigma$ Browsing positions ratios where two concepts appear	Reinforce or evaporate concepts weights depending whether they appear in every new session or not	(Yu et al., 2012)
	✓			Tags	No weights for nodes were used.	Co-occurrence techniques. If two tags are used in combination by a certain user for annotating a certain bookmark	Edges weights incremented by 1 each time they co-occur. Edges with top k weights will be selected to represent user profile	Edges weight evaporation by removing small percentage each time the profile is updated	(Michlmayr & Cayzer, 2007)

## 6. Conclusion

In this paper, we have provided an overview of the latest research efforts in the area of user modelling in social media websites. The user's profile representation and construction methods were discussed, and an ideation of the future work in this area has been provided. We have tried to cover all the possible options for the users' profiles design decisions, starting from the type of data that can be used and ending with the possibility of making dynamic profiles. In the future work section we highlighted some of the important aspects that can enhance the accuracy of users' profiles and overcome the weakness in the available models.

## References

- Abel, F., Gao, Q., Houben, G. J., & Tao, K. (2011a). Analyzing user modeling on twitter for personalized news recommendations. *User Modeling, Adaption and Personalization*, 6787, 1-12. [http://dx.doi.org/10.1007/978-3-642-22362-4\\_1](http://dx.doi.org/10.1007/978-3-642-22362-4_1)
- Abel, F., Gao, Q., Houben, G. J., & Tao, K. (2011b). Semantic enrichment of twitter posts for user profile construction on the social web. *The Semanic Web: Research and Applications*, 375-389.
- Abel, F., Gao, Q., Houben, G. J., & Tao, K. (2011c). Cross-system user modeling and personalization on the social web. *User Modeling and User-Adapted Interaction (UMUAI), Special Issue on Personalization in Social Web Systems*, 22(3), 1-42. [http://dx.doi.org/10.1007/978-3-642-22362-4\\_1](http://dx.doi.org/10.1007/978-3-642-22362-4_1)
- Ahlqvist, T., Bäck, A., Halonen, M., & Heinonen, S. (2008). Social Media Roadmaps. *Helsinki: Edita Prima Oy*.
- Ahmed, A., Low, Y., Aly, M., Josifovski, V., & Smola, A. J. (2011). *Scalable distributed inference of dynamic user interests for behavioral targeting*. Paper presented at the ACM Conference on Knowledge Discovery and Data Mining (KDD) (pp. 373-382). <http://doi.acm.org/10.1145/2020408.2020433>
- Anand, S. S., & Mobasher, B. (2005). Intelligent techniques for web personalization. *Intelligent Techniques for Web Personalization*, 1-36. [http://dx.doi.org/10.1007/11577935\\_1](http://dx.doi.org/10.1007/11577935_1)
- Barla, M. (2011). Towards social-based user modeling and personalization. *Information Sciences and Technologies Bulletin of the ACM Slovakia*, 3(1), 52-60.
- Bhuiyan, T., Xu, Y., Jøsang, A., Liang, H., & Cox, C. (2010). Developing trust networks based on user tagging information for recommendation making. *WISE'10 Proceedings of the 11th international conference on Web information systems engineering* (pp. 357-364). Springer.
- Chen, J., Nairn, R., Nelson, L., Bernstein, M., & Chi, E. H. (2010). *Short and tweet: experiments on recommending content from information streams*. Paper presented at the Proceedings of the 28th international conference on Human factors in computing systems (pp. 1185-1194).
- Das, A. S., Datar, M., Garg, A., & Rajaram, S. (2007). *Google news personalization: scalable online collaborative filtering*. Paper presented at the 16th International Conference on World Wide Web (pp. 271-280). <http://dx.doi.org/10.1145/1242572.1242610>
- Degemmis, M., Lops, P., & Semeraro, G. (2007). A content-collaborative recommender that exploits WordNet-based user profiles for neighborhood formation. *User Modeling and User-Adapted Interaction*, 17(3), 217-255. <http://dx.doi.org/10.1007/s11257-006-9023-4>
- De Meo, P., Quattrone, G., & Ursino, D. (2010). A query expansion and user profile enrichment approach to improve the performance of recommender systems operating on a folksonomy. *User Modeling and User-Adapted Interaction*, 20(1), 41-86. <http://dx.doi.org/10.1007/s11257-010-9072-6>
- De Pessemier, T., Deryckere, T., & Martens, L. (2009). *Context aware recommendations for user-generated content on a social network site*. Paper presented at the Seventh European Conference on European Interactive Television (pp. 133-136). <http://dx.doi.org/10.1145/1542084.1542108>
- Gao, M., Liu, K., & Wu, Z. (2010). Personalisation in web computing and informatics: Theories, techniques, applications, and future research. *Information Systems Frontiers*, 12(5), 607-629. <http://dx.doi.org/10.1007/s10796-009-9199-3>
- Gauch, S., Speretta, M., Chandramouli, A., & Micarelli, A. (2007). User profiles for personalized information access. In B. Peter, K. Alfred & N. Wolfgang (Eds.), *The adaptive web* (pp. 54-89). Springer-Verlag.
- Gorry, G. A., & Westbrook, R. A. (2011). Can you hear me now? Learning from customer stories. *Business horizons*, 54(6), 575-584. <http://dx.doi.org/10.1016/j.bushor.2011.08.002>

- Gueye, M., Abdessalem, T., & Naacke, H. (2012). Dynamic recommender system: using cluster-based biases to improve the accuracy of the predictions. *arXiv preprint arXiv:1212.0763*.
- Hanna, R., Rohm, A., & Crittenden, V. L. (2011). We're all connected: The power of the social media ecosystem. *Business horizons*, 54(3), 265-273. <http://dx.doi.org/10.1016/j.bushor.2011.01.007>
- Hannon, J., Bennett, M., & Smyth, B. (2010). *Recommending twitter users to follow using content and collaborative filtering approaches*. Paper presented at the Fourth ACM Conference on Recommender Systems (pp. 199-206). <http://dx.doi.org/10.1145/1864708.1864746>
- Hannon, J., Mccarthy, K., O'mahony, M. P., & Smyth, B. (2012). A multi-faceted user model for twitter. *User Modeling, Adaptation, and Personalization*, 303-309. [http://dx.doi.org/10.1007/978-3-642-31454-4\\_26](http://dx.doi.org/10.1007/978-3-642-31454-4_26)
- Hung, C. C., Huang, Y. C., Hsu, J. Y., & Wu, D. K. C. (2008). *Tag-Based user profiling for social media recommendation*. Paper presented at the Workshop on Intelligent Techniques for Web Personalization & Recommender Systems at AAAI2008, Chicago, Illinois.
- Jadhav, A., Purohit, H., Kapanipathi, P., Ananthram, P., Ranabahu, A., Nguyen, V., ... Sheth, A. (2010). Twitris 2.0: Semantically empowered system for understanding perceptions from social data. *Semantic Web Challenge*.
- Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons*, 53(1), 59-68. <http://dx.doi.org/10.1016/j.bushor.2009.09.003>
- Kietzmann, J. H., Hermkens, K., Mccarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business horizons*, 54(3), 241-251. <http://dx.doi.org/10.1016/j.bushor.2011.01.005>
- Kim, H. N., Ha, I., Lee, K. S., Jo, G. S., & El-Saddik, A. (2011). Collaborative user modeling for enhanced content filtering in recommender systems. *Decision Support Systems*, 51(4), 772-781. <http://dx.doi.org/10.1016/j.dss.2011.01.012>
- Li, L., Yang, Z., Wang, B., & Kitsuregawa, M. (2007). Dynamic adaptation strategies for long-term and short-term user profile to personalize search. *Advances in Data and Web Management*, 228-240. [http://dx.doi.org/10.1007/978-3-540-72524-4\\_26](http://dx.doi.org/10.1007/978-3-540-72524-4_26)
- Lops, P., De Gemmis, M., Semeraro, G., Musto, C., Narducci, F., & Bux, M. (2009). A semantic content-based recommender system integrating folksonomies for personalized access. *Web Personalization in Intelligent Environments*, 27-47. [http://dx.doi.org/10.1007/978-3-642-02794-9\\_2](http://dx.doi.org/10.1007/978-3-642-02794-9_2)
- Lu, C., Lam, W., & Zhang, Y. (2012). *Twitter User Modeling and Tweets Recommendation Based on Wikipedia Concept Graph*. Paper presented at the Twenty-Sixth Conference on Artificial Intelligence Workshops (AAAI).
- Ma, H., Zhou, D., Liu, C., Lyu, M. R., & King, I. (2011). *Recommender systems with social regularization*. Paper presented at the Fourth ACM International Conference on Web Search and Data Mining (pp. 287-296). <http://dx.doi.org/10.1145/1935826.1935877>
- Mezghani, M., Zayani, C. A., Amous, I., & Gargouri, F. (2012). *A user profile modelling using social annotations: a survey*. Paper presented at the Proceedings of the 21st international conference companion on World Wide Web, Lyon, France. <http://dx.doi.org/10.1145/2187980.2188230>
- Michlmayr, E. (2007). *Learning user profiles from tagging data and leveraging them for personal (ized) information access*. Paper presented at the S. Golder & F. Smadja (Chairs) Tagging and Metadata for Social Information Organization. Workshop held at the 16th International World Wide Web Conference. Retrieved December, 2008.
- Noble, C. H., Noble, S. M., & Adjei, M. T. (2012). Let them talk! Managing primary and extended online brand communities for success. *Business horizons*, 55(5), 475-483. <http://dx.doi.org/10.1016/j.bushor.2012.05.001>
- Noll, M. G., & Meinel, C. (2007). Web search personalization via social bookmarking and tagging. *The Semantic Web* (pp. 367-380). Springer.
- Riggs, T., & Wilensky, R. (2001). *An algorithm for automated rating of reviewers*. Paper presented at the 1st ACM/IEEE-CS Joint Conference on Digital Libraries (pp. 381-387). <http://dx.doi.org/10.1145/379437.379731>

- Saleh, I. (2012). Egypt's digital activism and the Dictator's Dilemma: An evaluation. *Telecommunications Policy*. <http://dx.doi.org/10.1016/j.telpol.2012.04.001>
- Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613-620. <http://dx.doi.org/10.1145/361219.361220>
- Santos Jr, E., & Nguyen, H. (2009). Modeling users for adaptive information retrieval by capturing user intent. *Collaborative and Social Information Retrieval and Access: Techniques for Improved User Modeling*. *IGI Global*, 88-118. <http://dx.doi.org/10.4018/978-1-60566-306-7.ch005>
- Semeraro, G., Degemmis, M., Lops, P., & Basile, P. (2007). *Combining learning and word sense disambiguation for intelligent user profiling*. Paper presented at the Proceedings of the 20th international joint conference on Artificial intelligence (pp. 2856-2861).
- Tao, K., Abel, F., Gao, Q., & Houben, G. J. (2012). *TUMS: twitter-based user modeling service*. Paper presented at the The Semantic Web Workshops (ESWC 2011) 269-283. [http://dx.doi.org/10.1007/978-3-642-25953-1\\_22](http://dx.doi.org/10.1007/978-3-642-25953-1_22)
- Tauscher, L., & Greenberg, S. (1997). How people revisit web pages: Empirical findings and implications for the design of history systems. *International Journal of Human-Computer Studies*, 47(1), 97-137. <http://dx.doi.org/10.1006/ijhc.1997.0125>
- Wall, T. V. (2007). Folksonomy. Retrieved from <http://vanderwal.net/folksonomy.html>
- Wen, J., & Zhou, W. (2012). An Improved Item-based Collaborative Filtering Algorithm Based on Clustering Method. *Journal of Computational Information Systems*, 8(2), 571-578.
- Weng, J., Lim, E. P., Jiang, J., & He, Q. (2010). *Twitterrank: finding topic-sensitive influential twitterers*. Paper presented at the Proceedings of the third ACM international conference on Web search and data mining (pp. 261-270). <http://dx.doi.org/10.1145/1718487.1718520>
- Xiang, L., Yuan, Q., Zhao, S., Chen, L., Zhang, X., Yang, Q., & Sun, J. (2010). *Temporal recommendation on graphs via long-and short-term preference fusion*. Paper presented at the Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 723-732). <http://dx.doi.org/10.1145/1835804.1835896>
- Xu, T., & Oard, D. W. (2011). Wikipedia-based topic clustering for microblogs. *Proceedings of the American Society for Information Science and Technology*, 48(1), 1-10. <http://dx.doi.org/10.1002/meet.2011.14504801186>
- Yang, X., Steck, H., & Liu, Y. (2012). *Circle-based recommendation in online social networks*. Paper presented at the 18th ACM International Conference on Knowledge Discovery and Data Mining (KDD). 1267-1275. <http://dx.doi.org/10.1145/2339530.2339728>
- Yu, J., Liu, F., & Zhao, H. (2012). *Building User Profile based on Concept and Relation for Web Personalized Services*. Paper presented at the International Conference on Innovation and Information Management.
- Yu, L., Pan, R., & Li, Z. (2011). *Adaptive social similarities for recommender systems*. Paper presented at the Proceedings of the fifth ACM conference on Recommender systems, Chicago, Illinois, USA. <http://dx.doi.org/10.1145/2043932.2043978>
- Zhong, E., Fan, W., Wang, J., Xiao, L., & Li, Y. (2012). *ComSoc: adaptive transfer of user behaviors over composite social network*. Paper presented at the 18th ACM International Conference on Knowledge Discovery and Data Mining (KDD)696-704.
- Zhou, X., Xu, Y., Li, Y., Jøsang, A., & Cox, C. (2012). The state-of-the-art in personalized recommender systems for social networking. *Artif. Intell. Rev.*, 37(2), 119-132. <http://dx.doi.org/10.1007/s10462-011-9222-1>

## Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).