IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 25, 2019, accepted September 23, 2019, date of publication October 1, 2019, date of current version October 17, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2944927

A Survey on Personalized News Recommendation Technology

MIAOMIAO LI AND LICHENG WANG
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China
Corresponding author: Licheng Wang (wanglc2012@126.com)

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61972050, and in part by the Shandong Provincial Key Research and Development Program of China under Grant 2018CXGC0701.

ABSTRACT In the face of massive data on the Internet, users often ''lost themselves''. Personalized recommendation technology has made breakthroughs in areas such as e-commerce, advertising, audio and video recommendation in recent years. Due to the inherent characteristics of network news, such as the massive data, heterogeneity, update and change fast, timeliness and strong geographical awareness and so on, the progress of personalized recommendation technology in the field of news lags behind the above areas. And it cannot meet the requirements in news field entirely. Therefore, it is the main task of the current news recommendation system to integrate the existing personalized recommendation technologies into the news recommendation field, to study how to handle massive heterogeneous news data, to construct an optimal user preference model, and to improve the overall performance of the personalized news recommendation. This paper presents the state-of-the-art personalized news recommendation technologies in recent years, and analyzes the advantages and disadvantages of the mainstream technology based on seven main directions. Finally, the open issues in the development of personalized news recommendation technology are analyzed and concluded, hoping to guide the related work in the future.

INDEX TERMS Personalized news recommendation, context-aware, data mining.

I. INTRODUCTION

With the rapid development and extensive applications of Internet information technology, the Internet has gradually become an important channel for people to access information. Hundreds of millions of network information show up in the world every day, and people gradually enter the information overload era from the information deficient age [1]. In the face of such a huge amount of information, Internet users often cannot be quickly and efficiently access to really valuable information they need. Personalized recommendation technology [45] is a tool to help users quickly find out the information they are most likely to be interested in. It can provide different users with personalized services to meet their specific interests and needs. In contrast to the other two tools used to address information overload problems, such as taxonomies and search engines, personalized recommendation technology does not require the users to provide explicit and exact requirements. Instead, it builds the users' interest preference model by analyzing their historical behavior records and other relevant information, then proactively recommend information that best meets their interests and needs to users based on the model.

Today, it is an important means for people to access information that they read news posted on the Internet now and then. A large number of news websites and apps provide people rich and abundant information sources to understand the world beyond their own world and shorten the distance with others. At the same time, massive news information also brings users new problems and challenges in finding interesting news. On the one hand, there are countless sources of news and massive news, making it difficult for users to make choices among vast amounts of news. On the other hand, different news websites and apps have different resources and backgrounds, leading to messy contents of news [4]. Personalized news recommendation technology [46] applies personalized recommendation technology in the field of news. It is committed to helping users quickly and efficiently access to news which fits users most from massive news information on the Internet, and mining potential interests of the users

The associate editor coordinating the review of this manuscript and approving it for publication was Junchi Yan.

to achieve personalized recommendation service for news readers. The benefits are that users do not have to spend too much time searching for news, which can save time and efforts as well as increasing users' satisfaction. What's more, News writers and news sites or apps maintainers also have greater financial benefits.

Different from the recommendation of items in the fields of e-commerce, tourism, movie, music and so on, the design and application of personalized news recommendation techniques are more complicated and difficult because of the characteristics of news itself [18], [20], [27], [28] (strong contextual correlation, rapid changes of popularity, strong timeliness performance, social impact factors, etc.) and the relevance between news (news is not independent). In addition, due to the huge amount of news information on the Internet, the large number of readers and their rapid growth, it brings a high demand on the massive data processing capability of personalized news recommendation technology. A good personalized news recommendation system must also have a strong adaptability and scalability, to provide users with a small response time delay in news recommendation services [16], [30], [36], [39], [43]. But such news recommendation systems with both high recommender effect and good performance are still difficult to entirely implement now by using the existing techniques. Thus, there are still aspects to be studied and improved in personalized news recommendation [46].

This paper mainly reviews the research status and progress of personalized news recommendation technology. Section 1 provides a brief introduction and overview of the basics of personalized recommendation techniques and personalized news recommendation techniques. Section 2 shows the overall framework and the main process of personalized news recommendation. Section 3 introduces data access methods and data processing techniques, which are the upfront work that personalized news recommendation must do. Section 4 focuses on analyzing some of key technologies in personalized news recommendation considering different news features. Section 5 points out the key issues and difficult issues in personalized news recommendation. Section 6 is a conclusion of this paper.

II. AN OVERVIEW ON PERSONALIZED NEWS RECOMMENDATION

News is one of the most important carriers of Internet information. Due to the information overload problem caused by the development of the network, it is more and more difficult for users to find news that they are really interested in when they face the massive news information. Therefore, the research on personalized news recommendation technology has drawn more and more attention from all walks of life. It is the main task in current personalized news recommendation research to design the algorithms with both high recommendation effect and good performance by intelligently combining the existing personalized recommendation technologies with the unique characteristics of news. In light of this, the next step is to outline the basics of personalized recommendation first, followed by personalized news recommendation.

A. PERSONALIZED RECOMMENDATION

At present, there are a lot of researches on personalized recommendation, and many mature recommendation methods are spawned. The most familiar methods are the following: recommendation based on association rules, collaborative filtering, content-based filtering, social filtering and hybrid recommendation.

1) RECOMMENDATION BASED ON ASSOCIATION RULES

Association rules mining refers to finding the interesting correlation or correlation between items in a large amount of data. The core idea is to discover rules and patterns and relationships between these patterns contained in data which meet certain support and confidence levels through mining the known data. Recommendation based on association rules [47] is to establish the matching relationships between items and user interest models by using the various association rules that are mined, and then predict users' interests based on these relationships, and then give recommendations.

This method is simple, direct and easy to implement. It also has a strong versatility in various fields and a better real-time performance in recommendation. However, there is a serious problem of item cold start [56]. The newly added item is hard to be found by the rules in the system and cannot be recommended due to the lack of user behavior data about it. With the increasing number of items and users in the system, rules grow rapidly. So the maintenance costs of the rules are correspondingly increased, which reduces the operating efficiency of the system [1].

2) COLLABORATIVE FILTERING RECOMMENDATION

Collaborative Filtering Recommendation [48] (CF in short) uses the preferences of a like-minded group with common experience to recommend interesting items to users. It finds relevance between items or users by analyzing users' behaviors, then accordingly recommends to users. CF recommendation is the earliest proposed and the most widely used method in recommender systems. Its basic idea is group intelligence. CF can usually be divided into two methods: User-Based Collaborative Filtering (User-Based CF in short) and Item-Based Collaborative Filtering (Item-Based CF in short).

a: USER-BASED CF

User-Based CF [48] finds similarities between users through user preference analysis on items, and then makes recommendations based on similar users. Specifically, it finds the user set with similar interest to the target user according to users' preference behavior data firstly, and then filters out those items which are preferred by similar users but not browsed by the target user in all item sets, and finally recommend those items to the target user. The recommendation accuracy

of User-Based CF is high in the case of a complete and rich dataset without excavating contents of items. Besides, User-Based CF can excavate the relevance between the recommended objects and the user preferences implicitly and transparently [48]. However, with the increase of the number of users in the system, this method needs to maintain a huge user similarity matrix, which makes the recommended calculation time longer and reduces the system efficiency [1]. CF does not mine the content of the item, and User-Based CF method cannot solve the item cold start problem either. New items will not be able to be recommended to the target users due to the lack of necessary user behavior data [56].

b: ITEM-BASED CF

Item-Based CF [49] use similar items for recommendation. It discovers similarities in the items, and then recommend similar items based on the users' existing choices to them. Specifically, Item-Based CF method calculates the similarity between items according to the user behavior data firstly, and then generates a recommendation list for the target user based on the similarity of the items and the user behavior. Item-Based CF recommendation method does not require the historical behavior of new users when faced with new users. Once new users act on one item, they can be recommended those items similar to it. At the same time, Item-Based CF is able to feedback users' behaviors quickly. That is to say, a user's new behavior will lead to real-time changes in the recommendation result. Compared with User-Based CF, it can make a good explanation of recommendation to users using historical behaviors of users, and the recommendation result is more convincing. However, it is precisely because this method uses the users' behaviors to explore the similarity between items without considering the interest differences of different users, nor does it take into account the content relevancy of items, the recommendation accuracy of Item-Based CF gets lower than that based on users. At the same time, it does not have the means to recommend new items to users without updating the item similarity table offline [56]. Especially, in the news recommendation field, news updates very fast, far exceeding the increasing rate of new users. While this method requires maintaining and updating a huge news similarity table, making it cost very much and inefficient to calculate and update the table [1].

3) CONTENT-BASED RECOMMENDATION

Content-Based Recommendation [50] (CB in short) is an extension and development of CF recommendation. This method excavates and analyzes the contents of the recommendation objects, obtains users' interests based on users' historical behaviors, and recommends to users the items that best match their interest models in the content. The core of this method lies in the mining of content features of recommendation objects and the construction of interest models based on content features and users' behaviors [27]. Information filtering method mentioned in [19] is a variant of CB essentially.

CB personalized recommendation method can generally be divided into three steps. Firstly, some features are extracted for each item based on the content to represent each item, called text representation of the recommendation objects. Then, users' favorites and interests are learned and constructed by using the features of item sets that users like or dislike in the past, called users' preference model construction. Finally, a set of most relevant items are selected to recommend for each user based on users' interests and features of candidate items obtained in the previous two steps, called generation of the recommendation results.

Though CB recommendation does not require the data of other users, it can still capture users' interests accurately. The recommendation effect of CB is more accurate, and the newly-appeared recommendation objects and non-popular objects can also be recommended in this method. It is able to solve the problems of cold start and sparseness in CF recommendation. But there are also some limitations in CB. For example, some rules between different objects cannot be learned by machine learning tools, or some content features of objects are very difficult to extract. Those objects, such as multimedia data, will not be effectively recommended. And, the over-characterization problem in recommendation process may lead to a shortcoming that the objects which users have never acted on will not be recommended. It suffers from a problem of pushing very similar recommendations to users. Thus, this method may lack novelty, resulting in users losing the opportunity to find different types of information.

4) SOCIAL FILTERING RECOMMENDATION

If an item is liked by other users who have intersections with the target user on the social network, the item may be recommended to the target user, which is called social filtering recommendation [52]. The advantage of this method is that it does not require interactions between users and items but still can make recommendations for new users or new items. That is to say, it is able to solve the cold-start problem. It is because that this method uses users' social network information to analyze their interest preferences, and then selectively recommend their friends' favorite items to them, which is somewhat similar to user-based CF method, except that the data sources are different. The disadvantage of this method is that data sparseness in social network still exists, and it does not take the contextual information into account.

5) HYBRID RECOMMENDATION

As the various recommendation methods developing, there comes a number of studies about mixing these multiple recommendation methods. Hybrid recommendation [53] mainly uses such mixing strategies [54]: weighted fusion, switch back and forth, feature combination, cascade, meta-level hybrid, feature augmentation, etc. to integrate varieties of recommendation methods. The purpose of hybrid recommendation is to make up for the shortcomings of single methods, to maximize their advantages themselves, and learn from each

TABLE 1. Characteristics of personalized news recommendation.

The Main Characteristics of Personalized News Recommendation	
Context-awareness	Users' preferences for news rely on the users' current contexts to a certain extent.
Social influencing factors	Famous people, friends or events on the social network can easily affect readers' interest.
Popularity effect	Users may be of great interest in the explosive and popular news on the Internet.
Timeliness	News are very time-sensitive, and update very fast. News have very short life cycles.
Massive data	The amount of news and readers are both huge, and the growth rate is very high.
Unstructured data	Users' interaction data and news texts are mostly unstructured.

other. In practice, people usually choose different recommendation methods and use appropriate strategies to mix them according to specific application scenarios.

B. PERSONALIZED NEWS RECOMMENDATION

News is a statement of recent or ongoing facts that are meaningful and interesting to the public [2]. As the recommendation object in personalized news recommendation, news is different from other objects to be recommended, such as e-commerce products, videos, music and so on. Therefore, compared with other personalized recommendation such as goods recommendation, movies recommendation, music recommendation, etc., personalized news recommendation has not only some similarities but also its own unique characteristics.

Generally speaking, apart from the common problems in all recommender systems such as sparseness, cold start and users' privacy, compared with other personalized recommendation, there are many unique problems in news recommendation, which are strong contextual awareness, many social influencing factors, fast update speed, a large amount of users, large differences in popularity, a high demand for processing speed, unstructured attribute of news texts, etc. So, higher requirements for the development of the unique personalized news recommendation technologies have been put forward due to these factors.

Specifically, compared with other personalized recommendation, there are mainly six characteristics listed in Table 1 in personalized news recommendation.

- Users' preferences for news are not only based on news topics and news content, but also based on the users' current contextual information such as user location, time, social information and major events at home and abroad. Therefore, personalized news recommendation needs to consider specific contextual information and relationships between different news.
- News readers are easy to be affected by their friends or events on the social network, so there are many social influencing factors that need to be considered in personalized news recommendation.
- Even though explosive and popular news are not related to users' interest preferences, they may also be of great interest to users due to users' herd mentality or other factors. Therefore, users' interest transfer should be fully taken into account.
- News have a strong timeliness performance, and update very fast. So each piece of news has a short life cycle. Making personalized news recommendation should focus on current news rather than outdated news.
- The amount of news data and the number of users are both huge, and the growth rate is very high. So personalized news recommendation technology must be able to adapt to such a large amount of data to provide scalable news recommendation service with a small response time.
- Users' interaction data and news texts are mostly unstructured in personalized news recommendation, and the unstructured attribute makes it more difficult to analyze the relationship between users and news.

Besides, the freshness factor of news may occupy higher weight than the correlation factor between news and users in some cases. And it is important for personalized news recommendation to recommend news with novelty to readers at any time. Since two irrelevant news texts may share lots of same or similar words, high similarity values based on words may not represent a strong relationship between news. And news readers may have special preferences for certain events in news, called named entity. Then Table 2 lists some differences between personalized news recommendation and other personalized recommendation on four main aspects.

From the comparison results showed in Table 2, we can get the following conclusion. Compared with several other personalized recommendations, personalized news recommendation has the strongest contextual awareness; it is easiest to be influenced by social factors; it is most time-sensitive; and it has a higher requirement on the scalability of systems due to the rapid growth rate of data. Therefore, using the existing recommendation technologies to implement a good news recommendation system with high performance is insufficient.

III. THE STRUCTURE OF PERSONALIZED NEWS RECOMMENDATION

The overall framework of personalized news recommendation is roughly the same as that of personalized recommendation, so there is no specific indication that the item in Figure 1 is news. The dashed box part in the right side is responsible for collecting various types of user characteristics, including user attribute characteristics and user behavior characteristics. User attribute features include demographic

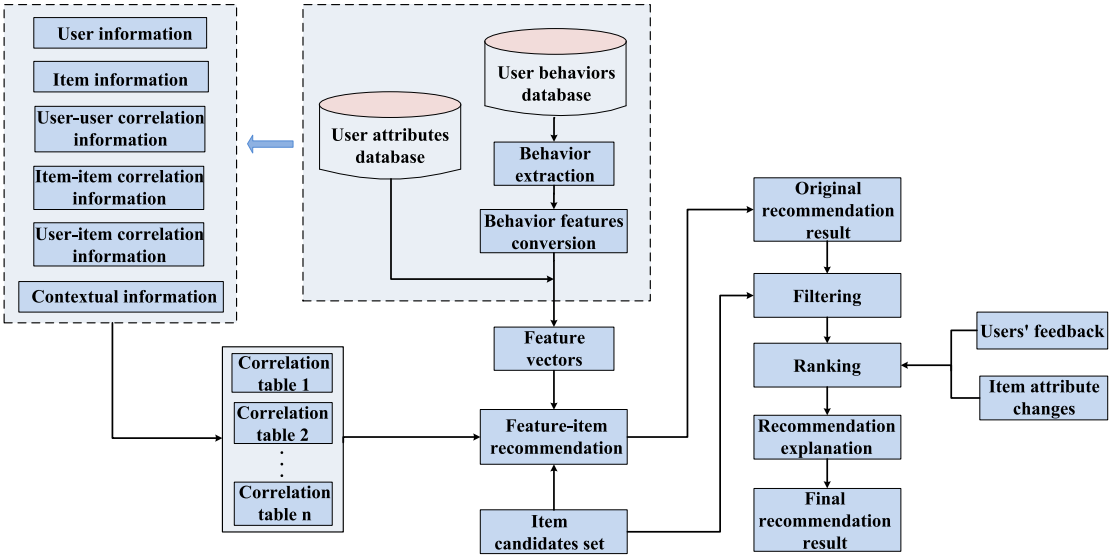

FIGURE 1. The overall framework of personalized news recommendation.

TABLE 2. A comparison between personalized news recommendation and other personalized recommendations.

	Context-awareness	Social factor influence	timeliness	scalability
E-commerce recommendation	weak	small	low	The amount of data increases rapidly, so it requires a high scalability.
Movie and video recommendation	weak	middle	low	general
Music recommendation	moderate	middle	low	general
Tourism recommendation	moderate	middle	medium	general
Book recommendation	weak	small	low	general
News recommendation	strong	big	high	The amount of data increases rapidly, so it requires a high scalability.

information such as age, gender, occupation, ethnicity, education, residence, etc. There are generally two types of user behavior characteristics: explicit feedback and implicit feedback. Explicit feedback refers to behaviors that users express their preferences for items explicitly, and the main ways to collect explicit feedback are rating and like or dislike option. Accordingly, implicit feedback refers to actions that do not clearly reflect users' preferences, such as page browsing behavior. Because browsing an item's page does not mean that the user like this item. Compared with explicit feedback, implicit feedback is more common, and it has a larger amount of data in practice.

The user information, user-user correlation information and user-item correlation information can be analyzed from users' various characteristics in the right dashed box part. These information, together with the item information, item-item correlation information and contextual information, can be used to generate the following correlation table based on some recommendation methods. Next, the initial recommendation list is generated from the set of candidate items based on the appropriate recommendation methods in combination with the extracted user feature vectors. Then, you can get the output (i.e. the final recommendation list) after the input (i.e. the initial recommendation list) goes through filtering, ranking and interpreting (sometimes) processes. The filtering module is based on the set of candidate items in order to eliminate items that the target user has acted on, non-candidate items and items in the blacklists, etc. The ranking module generates the optimal list of recommendation items according to the predetermined recommendation goal through a specific ranking algorithm. The ranking process is one of the most important steps of personalized recommendation. Particularly, it is more important in mobile recommendation scenario with limited screen size and high browsing cost. In addition, with the users' feedback and items' properties constantly changing, the ranking algorithms adjust from time to time.

In recent years, researchers have done a lot of work on ranking algorithms, such as content-based filtering, collaborative filtering and hybrid methods, etc. Most of these methods make recommendation ranking based on user profiles and the single-criterion approach, but users often make decisions based on multiple criteria. So there brings many systems beginning to use the multi-criteria approach [3]. And some researches [51] show that the use of multi-criteria

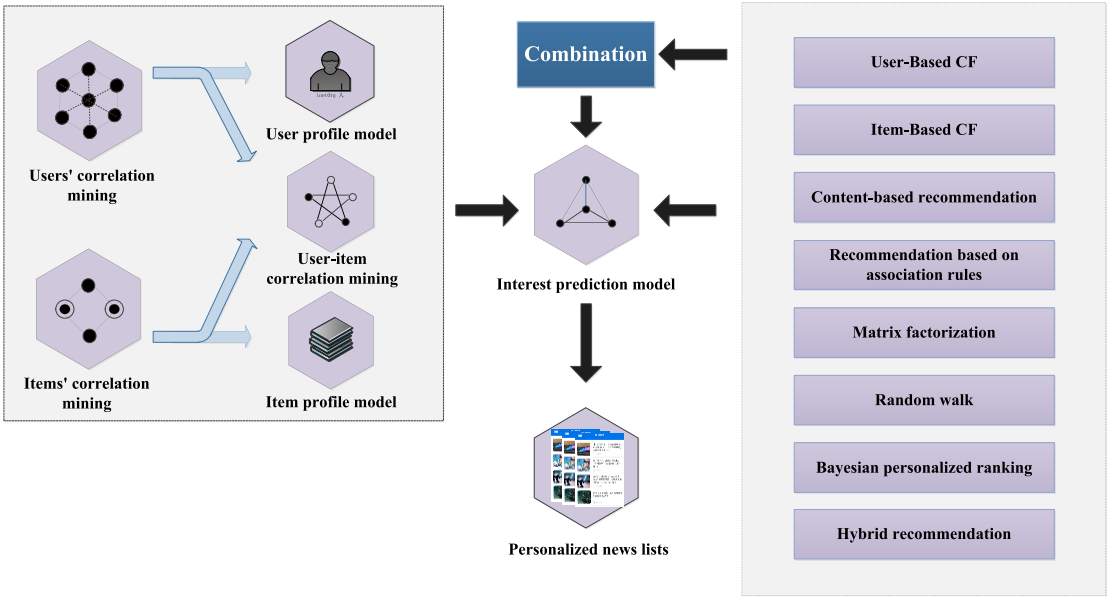

FIGURE 2. The main process of presonalized news recommendation.

approach makes it easier to integrate contextual information into the recommendation ranking methods. In addition, AHP-MCR [4] (Analytic Hierarchy Process based Multi-Criteria Ranking) is proposed to deal with the contextual information with dynamic real-time changes during the ranking process. It only gives a general AHP hierarchical model to support the function of adding or deleting contextual criteria or adjusting weights of different criteria flexibly corresponding to different application scenarios or user demands, rather than designing different AHP models for every different applications. In response to how to deal with recency and latency problems [60] in news ranking process, Moniz *et al.* [60] proposes an integrated framework that focuses on using resampling strategies to predict the importance of news, mainly the recent and popular news, and converting the predicted importance results into concrete news ranking results. Experiments in the paper show that this scheme well complements and enhances the news ranking results generated by existing recommender systems.

From the perspective of the current recommendation technology used in personalized news recommendation, the flow chart shown in Figure 2 can be drawn. By analyzing and mining user information and item information, one can get the user profile model, the item profile model and the user-item correlation model. Then a prediction model of users' interests can be extracted by training. And the news recommendation lists can be recommended to users who are most interested in the news based on the prediction model. What's more, the recommender system will show different lists for different users or user groups to achieve the final optative recommendation effect. The interest prediction model contains one or a combination of the eight recommendation algorithms listed in the right frame in Figure 2. Contextual information and social network information are often added to improve the recommendation accuracy and reduce the negative impact of data sparseness problem.

IV. DATA IN PERSONALIZED NEWS RECOMMENDATION

A. DATA SOURCES

Data is one of the core of recommendation system. For example, by analyzing data one can obtain the user preference models. Thus, dataset acquisition has always been the key issue for personalized recommendation systems and personalized news recommendation. Considering the privacy and security of Internet users, there are few datasets publicly available and effective online, which is a challenge for personalized news recommendation research. At present, there are mainly 6 data sources shown in Table 3 in personalized news recommendation research.

(1) Public dataset. Datasets published online are often used for scientific research and competitions, which are generally pre-processed or anonymized to ensure the privacy of users to a certain extent. Therefore, compared with datasets from other sources, it has a stronger authority, objectivity and persuasion. However, this kind of dataset is the least on the Internet. Currently the main way to access public datasets is to download them from websites that provide public datasets, which are often data-sharing websites that allow both downloading and uploading data or links. Infochimps [80] is this kind of website, with a large number of publicly available data resources and a wide range of categories. In addition, datasets can also be downloaded from some official websites, such as the Kaggle [81] contest dataset, which can be downloaded from the Kaggle official website. Currently, there are only a few public datasets which can be applied to recommendation system for analysis, and such datasets are usually

TABLE 3. Different kinds of datasets in presonalized news recommendation.

Dataset sources		Advantages	Shortcomings	Examples
Public dataset	Traditional news data	stronger authority, objectivity and persuasion; dependable; easy to use	few, limited application field, limited data scale	Kaggle news dataset, Kosarak[5]
	News data integrated with social information			Usenet Newsgroups[89], Twitter[34,35]
Crawling dataset from web crawler		strong objectivity and persuasion	existing "dirty" data and redundant data	RSS news data[4,10]
Dataset collected by investigation		authentic, dependable	small-scale dataset, subjectivity, hard to implement, cost much	Data of small-scale groups[11], Users data in Twitter[35]
Dataset provided by business company		objective, dependable, complete	hard to get	Mobile service's user data[12], Courses data of college students[39]
Derived dataset		flexible, easy to generate large-scale dataset	data reliability changes with rules	Dataset based on IMDb and MovieLens[14]
Simulated dataset		easy to generate large-scale dataset	lack of authenticity and credibility	MobileServices dataset[13]

targeted for specific recommendation fields. For example, last.fm [82] provides a dataset in music recommendation, Book-Crossing [83] is a dataset for book recommendation, and the famous MovieLens dataset [84] is for movie recommendation. Since there are two kinds of news, that is traditional news and news integrated with social information such as social network and social media, public datasets of news recommendation can also be divided into the following two categories.

- Traditional news data. For example, the competition dataset in the Kaggle contest includes news content recommendation and click-through data on Outbrain [85] website between June 14, 2016 and June 28, 2016. Kosarak [5], a clickstream dataset of online news websites, is from logs of the news web portals [86]. It only records users' news clickstream instead of recording excessive user click behaviors, and it is currently used in the tests of frequent itemset algorithms in association rules mining.
- News data integrated with social information. For example, the MicroBlog dataset [6] with a large size opened on the 2012 KDD Cup provides rich information in many fields such as user archives and social graphs. Usenet Newsgroups include user browsing data of various kinds of news groups. The contents of news groups and discussion topics contain computer-related technologies, entertainment, scientific or social discussions, arts, literature, politics and so on. Usenet users will rate and give feedback on these topics. Data in Twitter also contains rich user interest preference information, such as user id, users' tweets or re-tweets, time stamp, and topic tags, etc. The Friendfeed dataset [7] provides 1641531 posts from 111284 different users in 30 days from September 1, 2009 to September 30, 2009. Wikipedia [87] is an encyclopedia co-authored by Wikipedia users. Now it has been widely used in social network analysis and Wikipedia user behavior research [67]. For example, a user's editing-page behavior can be regarded as an implicit review of following this page with interest, and it is very helpful to analyze users' interest models for reading news by using Wikipedia thesaurus.

(2) Crawling dataset from web crawler. This method refers to grabbing website data randomly on the Internet to compose experimental dataset by writing crawler program or directly using crawler software. For example, currently popular crawlers written in Python can crawl data on MicroBlog, Zhihu and other websites. And Junar [89] is a good website for data capture and data delivery services. This kind of dataset is real, so it has a strong objectivity and persuasion. But the disadvantage is the existence of ''dirty'' data and data redundancy. It usually needs pre-processing to clean the data before using it. For example, Chen [8] uses crawler programs to grab news data from Google News as experimental dataset. Li *et al.* [9], [41] crawl news data from Yahoo! Today's module. Xie *et al.* [10] build an RSS monitor in order to collect real-time news on the Internet, which can keep track of news constantly from various news websites, such as CNN, ABC and USA Today. Once the monitor finds new news in the RSS file, the news text contents and images corresponding to the URL will be downloaded and stored to the local database. Then, MP3 files are generated as recommendation items in the personalized news radios—iNewsBox based on the corresponding news texts in the local database by using some TTS middleware. There are over 300 pieces of news downloaded to the local server every day by using this method.

Yeung *et al.* [4] also crawl the data of RSS news websites as experimental dataset.

(3) Dataset collected by investigation. Dataset collected by investigation refers to the collection of data obtained through investigation of the preferences or behaviors of specific user groups. The most common ways of user research are questionnaires and monitoring of specific user groups. Such datasets are highly truthful, mostly free of ''dirty'' data and redundant data, and can be used directly. But the size of such dataset is generally small and its application scenarios are limited. As it is difficult to fully guarantee the universality and objectivity of user groups during the selection process, the universality and objectivity of the dataset are not high. And, the cost is very high. Yang *et al.* [11] select a total of 136 undergraduates and postgraduates for questionnaires, and then the questionnaire results collected are used as the experimental dataset. Researchers in [35] obtain experimental datasets by monitoring the Twitter usage of 8 Twitter users within 3 months.

(4) Dataset provided by business company. Datasets from business company are usually provided for researches in specific fields, some of which are for the protection of user privacy and some involve commercial interests or even national security. Therefore, they are intended to be used exclusively for specific experiments and scenarios, not publicly. For example, the news dataset used to verify the performance of the proposed method of LOGO in [27] is such kind of dataset. It is the backend log provided by a commercial company that involves commercial secrets, and it is only used for the experiments in the paper. Besides, usage logs of 200 registered users within 3 months from March 2005 to May 2005 provided by a Korean mobile service provider are used in [12], which contain services information of news, movies and restaurants. It is not open to the public, either.

(5) Derived dataset and simulated dataset. Derived dataset refers to the dataset which is formed through data collection, supplement and integration processes based on several existing public datasets under some reasonable rules specified. As the rules can be freely adjusted according to the actual situation, such datasets are more flexible and convenient to generate large-scale datasets. Simulated dataset refers to the dataset that is automatically generated by certain reasonable rules or algorithms. It is also convenient for acquiring large-scale datasets. However, due to the lack of universality and authenticity of rules formulated, both datasets are easy to be subjective. As a result, the persuasiveness and credibility of these two kinds of dataset are not high enough. Datasets based on IMDb and MovieLens in [14] are derived datasets, and MobileServices datasets in [13] are simulated datasets.

Using different kinds of datasets in combination in experiments can improve the reliability and credibility of recommendation results to a certain extent. For example, both simulated dataset and questionnaire data are used in [11]. Similarly, both the investigation dataset and the derived dataset are used in [14].

TABLE 4. Indicators of some public datasets [90].

Dataset	Number of users	Number of items	Number of rating data	Rating density (%)	Rating range
Kosarak	-	41270	990002	8.1	counting
Movielens 1M	6040	3883	1000209	4.26	1～5
Last.fm	1892	17632	92834	0.28	counting
Book-Crossing	92107	271379	1031175	0.0041	1～10 and implicit behaviors
Wikipedia	5583724	4936761	417996366	0.0015	interaction
Jester[92]	124113	150	5865235	31.50	-10～10

Indicators of the public datasets listed in Table 4 contain number of users, number of items, rating density and the type of rating, etc. The rating density refers to the ratio of the average number of items rated by each user and the total number of items in the dataset. If every user rates every item, the rating density is 100%. Conversely, if each user does not rate each item, the rating density is 0. In fact, most of the ratings data are from only a small number of users, and the rating behaviors of the remaining users are very little. In practice, researchers can select the experimental datasets effectively by viewing these indicators. In addition, these public datasets also provide additional information about users and items, allowing researchers to explore effective ways to get user preferences from the datasets.

B. DATA PROCESSING

With data sources, we can get the initial data used in personalized news recommendation, which mainly includes user information, news information, contextual information, social media information and so on. However, these initial data may have dirty data, redundant data, data missing and other problems. Therefore, in order to normalize the initial data, it needs to be pre-processed. The pre-processing operation mainly includes the processes of checking, filtering, quantifying, structuring and filling, etc. Checking and filtering are designed to delete duplicate data and synonym data to reduce the amount of data. Quantitative representation is an effective way to convert unstructured data into structured data. For example, a user's click-browsing behavior on news can be quantified by binary representation. Besides, time context factor can also be quantified as different periods which are represented by unique values. Due to the small amount of explicit data on the Internet, a large number of implicit data are used to predict and fill in the missing data. Unstructured initial news data is inconvenient to the processing and calculation subsequently, so the unstructured news texts need to be

presented in a unified and structured way in order to facilitate the storage and processing afterwards.

At present, most commonly used methods of news texts representation come from the text mining field, such as Probabilistic Model [15], Vector Space Model [58], [59] and so on. Probabilistic Model method refers to obtaining the latent semantic structures and probability distributions of news texts by using different probability generation models. The most widely used probability generation models recently are LDA (Latent Dirichlet Allocation) [16], PLSA (Probabilistic Latent Semantic Analysis) [15], ESA (Explicit Semantic Analysis) [17] and so on. The PLSA model is relatively stable, which can reduce the dimensionality of high-dimensional vectors. But it cannot handle new texts, and it is prone to overfitting. LDA overcomes some deficiencies of PLSA. It is suitable for large-scale news groups, but it lacks the ability to model the relevance of news topics. ESA is a new semantic analysis technology, which uses knowledge-intensive Wikipedia information and the identifiable concept to represent news texts. It helps improving the interpretability, but the disadvantage is that it is complex to establish the knowledge warehouse in the preliminary. The VSM method is a method of expressing news texts by using a vector space formed by a set of terms and the associated weights. In this model, each news text can be represented by a tuple of $\{T, W\}$. T is a collection of all the terms in the news text. The terms including words, topics, named entities and so on, are generally obtained by removing participles and stop words from texts and semantic analysis. Commonly used word segmentation tools in text mining are CKIP [18], ECScanner [8] and so on. These tools can also apply to news texts. W is the weights corresponding to the terms. The commonly used methods of term weights calculation in VSM include frequency statistics [57], TF-IDF (Term Frequency-Inverse Document Frequency) [4] and a series of improved methods, among which TF-IDF method is the most widely used. With the development of text mining technology, the structured representation of news texts is also getting more and more mature, which is conducive to the further development of personalized news recommendation.

Data processing is a very important step both in personalized news recommendation and other personalized recommendations because the initial data retrieved from most data sources have more or less problems. If without data preprocessing, it will bring a very big burden on the similarity sets calculation, recommendation generation stage subsequently, and eventually affects the final recommendation. Especially, the negative impact gets more serious in the case of large amount of data.

V. KEY TECHNOLOGIES IN PERSONALIZED NEWS RECOMMENDATION

With personalized recommendation technology shining in the field of e-commerce, it has also gradually extended to other applications, including personalized news recommendation we discuss in this paper. However, as news has its own unique characteristics, which have been discussed in section 1.2, the key technologies of personalized news recommendation are a little different from that in traditional e-commerce field. Thus, the following parts bellow will give a comparison and analysis of the key technologies in personalized news recommendation from different perspectives of news characteristics.

A. USER PREFERENCES EXTRACTION FOR NEWS

In the research of personalized news recommendation, it is a very important process to accurately obtain user preferences for news, which directly affects the quality of interests matching and lists ranking subsequently, and eventually affects the final recommendation. However, it is very hard to obtain user preferences only by some simple methods, such as the click-through rate observation. Because the false click-through behaviors occupy a non-negligible proportion of the total click-through behaviors, obviously there are some deviations in user interest models calculated when regarding the false click-through behaviors as the click-through behaviors out of interests. Besides, there is very little explicit information that can indicate that a user really wants to read the news, which causes a great difficulty for user preferences acquisition. Therefore, researchers have come up with various combinations of multiple methods recently. Tavakolifard *et al.* [19] propose that users' demands can be analyzed by observing the ''pre-read'' action and the ''post-read'' action. The ''pre-read'' refers to the action of the user clicking on the news headline before reading the news, while the ''post-read'' represents actions of the user discarding, collecting, sharing, commenting, re-posting news links and other actions after reading the news. By combining ''pre-read'' actions with ''post-read'' actions together, researchers can analyze users' preferences fast and precisely, and then make better news recommendation to them. Wu *et al.* [20] propose the reading time-consuming factor α to correct the traditional user rating calculation formula in order to distinguish the users who are really interested in the news and the users who give a rough glance of the news, that is, to distinguish between clicks out of interests and false clicks. The experimental results indicate that this method can improve the recommendation accuracy to a certain extent. But the disadvantage is that it does not consider the external factors when users read news. For example, users interrupt reading news for some reasons and then continue to read the news after a period of time. So it is supposed to make precise adjustments on data collection methods, and refine the calculation of α at the same time in order to further improve the system performance.

B. DIFFERENT TECHNOLOGIES FOR DIFFERENT NEWS CHARACTERISTICS

1) TECHNOLOGIES FOR DATA SPARSITY AND COLD START

So far, researchers have tried many personalized recommendation technologies in personalized news recommendation field, such as the content-based approach, the collaborative

filtering method and combinations of multiple approaches and so on. Although they have made great progress by using the methods above, there are still two basic problems to be solved in personalized news recommendation. One is the data sparsity problem. Compared with the entire news set on the Internet, the data size of online users' news reading records is very limited. And the user-news interaction matrix shows a very strong sparseness, so it is difficult to obtain the similarities of different user reading modes effectively. The other is the cold start problem, which contains user cold start problem and item cold start problem. The former is due to the continuous changes of the user sets online, that is to say, there are constantly new users to join in and old users to exit or log out. The latter is due to the dynamic nature of the news itself.

Researchers have also done a lot of work on solving data sparsity and cold start issues. Studies in [21] indicate that the collaborative filtering method performs better than other methods when there is enough user rating data for news, but when the data is not enough, the collaborative filtering method cannot extract similar user groups. Therefore, a method integrated with real social network information is proposed in [22]. However, such real social network information is often not easy to obtain. Then, some researchers propose to model the information flow pattern in the virtual social network [23], in which people always influence each other inadvertently. Due to the absence of historical rating information of new users and new news, the previous collaborative filtering and social network approaches do not apply to both user cold start and item cold start issues any more. For new news, it is helpful to recommend it to the suitable users through analyzing its textual content [24]. For new users, a simple questionnaire can be posted on the news website, then the simple user profiles of new users can be constructed with their feedback in the questionnaire. The disadvantage is that it requires user explicit input behavior, and the quality of the questionnaire cannot be guaranteed either. Lin *et al.* [25] propose PRemiSE, which combines content information in virtual social networks, collaborative filtering and information diffusion technology into probabilistic matrix decomposition. In this way, semantic features of news and implicit user network structures are taken into consideration. And then with the guide of latent ''experts'', the cold start problem may be solved. Yeung *et al.* [4] propose a strategy using Bayesian network to eliminate the data sparseness problem for new users, that is, the user cold start problem. A Bayesian network is a probabilistic graphical model which combines the advantages of CF and CB. And three phases (i.e. construction, prediction and revision) are designed to resolve the cold start problem in [4]. A hybrid recommendation method is used in [19], which takes into account many recommendation factors such as news freshness, news popularity, users' long-term and short-term interests and users' contextual information. For new users, they can be recommended the latest or hottest news. For new items, there is a good chance of being recommended to many users given their high freshness values. Thus, there are no cold start issues with the system.

2) TECHNOLOGIES FOR RICH CONTEXTUAL INFORMATION

News is used for reporting the recent, nearby and valuable facts to people quickly and timely, with concise texts [92]. The ''recent'' factor refers to the time contextual information of news, and the ''nearby'' factor refers to the location contextual information of news. These two factors are the main contextual information considered in personalized news recommendation. In addition, there are many other contextual factors such as weather, season, weekday or weekend, crowd density and whether there are peers or not [26]. Faced with so much contextual information, Baltrunas *et al.* [26] propose a method to evaluate and model the relationship between contextual influential factors and item rating, and they also devise a strategy that inquires users about whether the particular contexts influence their decision making or not. The goal is to determine which types of contextual factors are more important and how much they affect the ratings. But we only discuss the impacts of time and location contextual information on personalized news recommendation in this paper. With the popularization of portable mobile devices such as cell phone, iPad and laptop, news readers are gradually shifting from the web side to the mobile side. Due to the portability and flexibility of mobile devices, users can read news anywhere at any time. Therefore, there is an urgent need to study the personalized news recommendation technology that integrates time and location contextual information.

The goal of personalized news recommendation incorporated with time and location information is to predict the users' reading interests at a specific time point (or period) and a specific place accurately. Taking the TopN recommendation as an example, the goal is to generate a list of news of length N for the users eventually, and the list should contain news that the users are most likely to be interested in at a certain moment or place. Detailed analyses on news recommendation technology with time factor and location factor are as follows, respectively.

a: TIME FACTOR

The impact of time contextual information on personalized news recommender system and user interests is extensive and deep, and is mainly reflected in three aspects. Firstly, user interests change constantly. Due to users themselves, their interests may change gradually or abruptly over time [44]. In order to extract the current interest preferences of users with gradual interest changes accurately, their recent behaviors should be concerned about, as recent behaviors can best reflect the current interest preferences. Secondly, news has a life cycle. The life cycle of news is short. Hot news today may have no hit tomorrow. Therefore, it is necessary to discriminate whether the news is out of date or not at a certain moment when considering recommending news to users at that moment. Thirdly, news has the time effect. For example, different seasons and festivals will influence users' interests, which is an effect caused by time changing. In general, the time factor needs to be considered when designing the recommendation algorithms.

In the face of the case that users' interests change over time, a new recommendation method is proposed in [27], which divides users' interests into two types—long-term reading preference and short-term reading preference. Long-term reading preference is obtained based on a design of time-sensitive weighting, while short-term reading preference is obtained by analyzing users' most recent reading history. Then different news groups preferred by different users can be separated by using users' long-term reading preferences, and then using users' short-term reading preferences this method is capable to select news which are most likely to be preferred by users from their corresponding news groups to recommend. This method combines the user's long-term interests with short-term interests completely, taking full account of the characteristic of users' interests changing gradually over time. Wu *et al.* [20] consider the length of reading time, which is the time contextual information of users reading news, and put forward the concept of reading time-consuming factor α accordingly. α represents the length of time users spend reading a piece of news. Using it researchers can distinguish between users who read news carefully and users who read news roughly. Correcting the traditional user rating formula by using α can improve the accuracy of recommendation to a certain extent. However, this method fails to consider the external influential factors when users read news. For example, the length of reading time is too long sometimes, as users may be interrupted when they are reading news, and then continue to read the news after a period of time. In the study of analyzing news time sequence information, Garcin *et al.* [28] propose a recommendation algorithm based on Context Trees (CT in short), which are updated fully incrementally. A CT is defined as a partition tree that is organized in a hierarchy of increasingly precise partitions of a space of contexts. The contextual information includes news sequence information, topic sequence information and topic distribution. With the core idea of giving recommendation results as contexts gradually become fine-grained and deep-hierarchical in the partition tree. This approach performs well in prediction accuracy and recommendation novelty aspects.

In personalized news recommendation, the temporal context plays a very significant role in the User-Based CF recommendation method, especially for the similarity measure. After finding a similar user group for a user u, the recent reading interest of the users in this group is apparently closer to the current reading interest of user u than their reading interest a long time ago. That is to say, the recent favorite news of users who have similar interests with the target user should be paid more attention when recommending news to the target user. It is especially important in the field of news recommendation with very short life cycles. In view of this, researchers have made the following improvement to the user similarity calculation formula [1].

$$sim_{uv} = \frac{\sum_{i \in N(u) \cap N(v)} \frac{1}{1+\lambda |t_{ui} - t_{vi}|}}{\sqrt{|N(u)|\,|N(v)|}} \tag{1}$$

In Eq. (1), a time decay factor on news i shared by user$u$ and user v is added to the numerator of this formula. $N(u)$ or $N(v)$ means a collection of news that user u or user $v$has read. t_{ui} represents the time point of u reading i. λ is a coefficient. The farther the time points of u and v reading news i, the smaller their similarity is. As a result, the similar groups of different users to be found will be more accurate.

b: LOCATION FACTOR

The personalized news recommendation considering the location information mainly considers the representation and retrieval of user location trajectory and the construction and recognition of the location trajectory network. Specifically, the location points of the user changing over time are connected in series to form a position movement trajectory based on the behavior logs of the user location movements, and then extracting the movement rules of the user position from the trajectory to construct all users' position movement trajectory network, and finally analyzing all users' different position movement modes based on the trajectory network in order to formulate a new neighbor user calculation method. When only considering a single user's position movement information, the future position of the user can be predicted according to the extracted user movement rule. Then the user will be recommended news which are most likely to be of interest in the range of the predicted position so as to realize the location-based personalized news recommendation.

At present, there are many researches on personalized news recommendation based on location information, which have already achieved good progress. Bao *et al.* [29] propose an improved location-aware news feed system. The system adopts three methods of news feed, namely spatial pull, spacial push and shared push, which can improve the shortcoming of the existing similarity calculation methods which consider only static point position and extend the position to a spacial scope. Soon after, this research team proposes another framework of location-aware news recommendation against mobile users [30]. There are mainly three modules in this framework: location prediction, similarity measurement and news recommendation generation. The location prediction module is to predict the future position of the mobile users based on the existing path prediction algorithm; the similarity measurement module combines the spatial position factor with the original vector space model to calculate the user-news correlation jointly; the news recommendation generation module generates the recommendation list to the user based on the current position and the predicted position of the user together. This framework is with high efficiency. Yin *et al.* [16] propose a location-aware recommendation system which adopts a method of modeling offline and recommending online considering user interest preference and geographical preference synthetically. Specifically, the user preference model and geographical preference model are learned offline. And then TopN recommendation is performed online according to the matching status of the user interests and

the location information. Experiments show that the system performs better than other systems in terms of efficiency and effectiveness, but the accuracy is not significantly improved. Son *et al.* [17] propose an improved ELSA (Explicit Localized Semantic Analysis) algorithm in order to make up for the shortcoming that ESA (Explicit Semantic Analysis) algorithm is insensitive to the location context. It makes use of the topic model LDA integrated with the location information to construct the news topic distribution, and then calculates the similarity based on the topic distribution, and finally makes recommendations. Experiments show that ELSA is superior to LDA, ESA and PESA (Probabilistic ESA) on the NDCG@k indicator. At present, there are many researches on location-aware personalized news recommendation technology, which is also an important direction of context-aware news recommendation system.

So far, there are still two problems in the research of context-aware recommendation systems. One problem is that each recommendation system is basically designed for a particular application scenario, thus limiting the scalability of the recommendation system. The other problem is that the context-aware recommendation systems all contain similar contextual information, but the same contextual information may be in different structures in different systems, which limits the data-sharing among different context-aware recommendation services. In response to the problems, Yeung *et al.* [4] propose a hybrid peer-to-peer context-aware framework JHPeer that supports a variety of context-aware services in the mobile environment, including news feeds. And all context-aware applications based on the JHPeer framework can reuse the contextual information collected by JHPeer. This framework can greatly expand the scalability. Studying the scalable context-aware recommendation techniques is of great importance for the promotion in practical applications.

3) TECHNOLOGIES FOR SOCIAL INFORMATION

In the 19th century, the German sociologist Ferdinand Tönnies considered social groups to be divided into two kinds [1]. One is formed through the common interests and beliefs of people, named ''community''. The other is formed due to the kinship and working relationship between people, called ''society''.

Social networks, such as Facebook and Twitter, allow users to create a public page that introduces themselves, and expose their buddy lists by default. Of course, users can also specify not to open certain buddies. And users rarely talk about topics involved with the personal privacy on the social networks, most of which are social hot spots or favorite music, videos, pictures and so on. Social networking websites alleviate the problem of information overload naturally because users tend to filter out information through their friends in the social network naturally when they navigate the web pages. Therefore, the personalized news recommendation system can make use of the public social network information and user behavior data in social networking websites to assist users to better complete the information filtering task, and to discover the user interest preferences faster, and finally making relevant news recommendation.

As the two representatives of social networking websites, Facebook and Twitter are actually two different types of social network structure. Users' friends in Facebook are generally people he knows in real life, and establishing a friend relationship needs confirmations by both of the two sides. While users' friends in Twitter are mostly people they do not know in reality. The friend relationships are usually established just because they are interested in each other's expressions of one's opinion or thoughts, and the friendships are generally one-way. Note that MicroBlog and Blog essentially belong to this type. The social network represented by Facebook is called the ''social graph'', and the social network represented by Twitter is called the ''interest graph''. Therefore, ''community'' and ''society'' that Ferdinand Tönnies says are the map of ''interest graph'' and ''social graph'', respectively.

a: INTEREST GRAPH

So far, there have been many studies on the use of Twitter data for news recommendation. More than 80% of the topics discussed in Twitter are related to news [31]. Studies in [32] show that the news on Twitter website spreads much faster than the traditional news. What's more, researches in [33] indicate that earthquake and typhoon forecasts can be made by analyzing news spread from Twitter. In summary, it shows that it is feasible and reliable to analyze and construct user preferences using Twitter information for news recommendation. Therefore, Phelan *et al.* [34] and Lee *et al.* [35] build users' personalized profiles based on users using Twitter status, that is, users' tweets, re-tweets and topic labels tagged, and then recommend the matching personalized news to users according to their profiles. The difference between [34] and [35] is that the former only uses the TF-IDF method to extract key words to build profiles, while the latter also introduces the LDA method to create topic models for news based on the former method, and the latter recommendation performance is advanced. As to MicroBlog, the relationships between users on MicroBlog are mostly one-way concern relationship which is called ''weak relationship'', relative to friends relationship in social graph—''strong relationship'', so the data on MicroBlog are more sparse, and it is difficult to give good recommendations only with one single CF algorithm or CB recommendation algorithm. Chen *et al.* [36] propose a hybrid two-phase clustering recommendation algorithm GCCR (Graph-Content Clustering Recommendation), which combines the graph abstracting clustering algorithm with the content similarity-based clustering method to filter out and extract dense datasets layer by layer. This method improves the recommendation accuracy greatly and ensures a certain degree of recommendation diversity at the same time. The disadvantages are the high complexity of the algorithm and the low implementation efficiency.

b: SOCIAL GRAPH

According to a survey done by Facebook's data scientist Lars Backstrom [93], 92% of Facebook's new friendships come from friends of friends, called FoFs. Therefore, for the social-graph social networks typified by Facebook, more than 90% of a user's online buddies are also real-life good friends. Considering the severe data sparseness in personalized news recommendation, we can see the important usage of such social network data in personalized news recommendation, especially in the similar user groups discovery process of the User-Based CF recommendation algorithm. In addition, such social network data can be imported into the personalized news recommendation system by using users' Facebook accounts to log in to the system, so that the data can be used to quickly cluster similar user groups with users' authorization and to help solving the user cold start problem, as well as improving the recommendation speed.

4) TECHNOLOGIES FOR POPULARITY EFFECT

One of the most important differences between news and other items is that news is extremely prone to be dominated by the incidents with high prevalence, that is, the so-called popularity effect. This is determined by the nature of news, which is intended to inform users of the latest and most popular information in the community. And news readers also would like to pay attention to the popular news. These popular news will be considered as the resources for online public opinion or the entertainments shared by personal friends. Even if the topics of the hot news do not match with the user interest model, the user will still be attracted over, which is the role of people's herd mentality.

Currently, the popularity of news both has a positive impact as well as a negative impact on personalized news recommendation. The positive impact is to guide the recommendation algorithm to consider allocating a certain weight to the current hot news (even if not matching with user interests) as well as requiring that news need to match with the user interest model. Generally, the hot news are news increasing rapidly in reading volume in a very short period. Such a comprehensive consideration of a variety of factors can better improve the user viscosity and satisfaction. Different news websites use news popularity differently to optimize algorithms in recommendation performance. Some only select the news whose popularity is greater than a given threshold as the candidate set for calculating the similar groups in the beginning. And some will give the popularity of the news a certain weight to participate the ranking process in the final ranking stage. For example, the Bayesian Personalized Ranking [42] (BPR in short) algorithm and its improved algorithm—HLBPR (Hybrid Local BPR) [37] are integrated into the hot spot factor, and finally display the optimal recommendation list.

Negative impacts can be mainly attributed to two types. One is that the similarities between popular news and other news are generally high. Take the situation of two users reading the same news at the same time for example. If there are few people reading this piece of news which may be cold news or professional news, the two users can be regarded as similar users then. But if the news has occupied the headlines of the major websites firstly, then the two users read it at the same time, it is not realistic to still consider them as similar users at this scenario. Therefore, there is a need to punish the hot news to mitigate this negative impact. And there are two major penalties to correct the similarity calculation method [18]. Note that the original calculation formulas for users' similarity and items' similarity are Eq. (2) and Eq. (3) respectively.

$$sim_{uv} = \frac{|N(u) \cap N(v)|}{\sqrt{|N(u)|\,|N(v)|}} \tag{2}$$

$$sim_{ij} = \frac{|N(i) \cap N(j)|}{\sqrt{|N(i)|\,|N(j)|}} \tag{3}$$

- Letting the popularity of the news i be $|N(i)|$, the value of $N(i)$ is determined by the size of the userset who read i, usually directly taking the size of the userset as $N(i)$. The user similarity calculation formula with the hot spot penalty factor becomes Eq. (4), where the term of $1/\log(1 + |N(i)|)$ is the hot spot penalty factor, used for amending the impact of common hot news in the user interest lists.

$$sim_{uv} = \frac{\sum_{i \in N(u) \cap N(v)} \frac{1}{\log(1+|N(i)|)}}{\sqrt{|N(u)|\,|N(v)|}} \tag{4}$$

- The other penalty is to correct the denominator of the original item similarity calculation formula, which turns into Eq. (5). $N(i)(N(j))$ means a collection of users who have read news $i(j)$. The revision method is to change the value of β. The hot news j will be punished by increasing the value of β. But this amendment sacrifices some of the recommendation accuracy.

$$sim_{ij} = \frac{|N(i) \cap N(j)|}{|N(i)|^{1-\beta}|N(j)|^{\beta}} \tag{5}$$

The other negative impact of news popularity factor is that hot news in different news fields also have relatively high similarities after identifying different news fields. It means that even if a user u does not like the news in domain I, the system will still recommend the hot news in I to u. Because the hot news in I have high similarity with the hot news in domain J which is favored by u. It is obviously unreasonable in this situation. At present, only using users' behavior data cannot solve this problem, so content information needs to be introduced, such as allocating different weights on different news fields, etc.

5) TECHNOLOGIES FOR TIMELINESS AND REAL-TIME EFFECT

For the personalized recommendation system, the timeliness of the system is closely related to the life cycle of the item, and the life cycles of the news are usually very short. The timeliness of the system can be measured by two indicators.

The first one is the average number of online days. The definition of an item online is that the item is acted on by at least one user within a day. Therefore, it is feasible to measure the life cycle of an item with the item average

number of online days, and then to measure the timeliness of such personalized recommendation systems. A test [9] shows that for different categories of items with the same average popularity, their average online days vary. For example, the news average number of online days in New York Times is short, while the words average number of online days in Wikipedia is very long, when they have the same level of popularity. It indicates that the timeliness of the two websites varies greatly. News websites like the New York Times are highly time-sensitive, because the life cycle of news is short, and popular news can only last for hours, then cool down quickly and be hit by the next wave of hot news. But words in Wikipedia website are different from news in nature, as users may search for certain words from time to time, and so the life cycles of words are much longer than news.

The second one is the average similarity of popularity vectors of items N days apart. Specifically, the item popularity of two days separated for N days in a website can be expressed by two vectors, then the average similarity calculated between the two vectors is considered as this indicator. (The average process refers to taking a multiple of two days apart from N days, then calculating similarities, finally taking average value). If the similarity value is big, it shows that there is not much difference between popular items apart from N days, that is, the items have long life cycles and a long average online time. Thus, the system is less time-sensitive. On the contrary, if the similarity value is small, it indicates that popular items in the system vary greatly in the interval of N days. Their life cycles are very short, and the average online time is short too. Thus, the system is much more time-sensitive. For example, for the New York Times and Wikipedia websites, although the average similarities of their N-day-apart item popularity vectors both decrease with the increase of N, the changing rates are different. The similarity descent rate of New York Times is large, while Wikipedia's similarity descent rate is relatively small. It indicates that the former changes rapidly in the popularity distribution of items while the latter changes slowly. And the former system is very time-sensitive while the latter's timeliness is relatively weak.

As interests of users have been constantly changing, a user's short-term interest may be contrary to his long-term interest, which mainly reflects in his new behaviors. So there is a certain requirement to the real-time effect of recommendation systems. A real-time recommendation system should meet the ever-changing interests of users and respond to new user behaviors quickly. Then it can ensure that the recommendation list changes with users' interests changing. A real-time recommendation system requires being able to access the user behaviors in real time. It also requires that the system recommendation result is calculated in real time according to the recent user behaviors around the visiting time point. And it requires to consider the user's long-term behavior and short-term behavior in a balanced manner. That is, short-term interest changes caused by users' short-term behaviors should be reflected out, as well as ensuring the continuation of predicting users' interests based on the long-term behaviors in recommendation lists. Generally, the requirements on the real-time performance of personalized news recommendation are basically the same as that in personalized recommendation system.

6) TECHNOLOGIES FOR MASSIVE DATA PROCESSING

With the massive growth of Internet news all the time and the huge number of users already existing, the research of personalized news recommendation has to consider the problem of data processing speed. The requirement on highly efficient and scalable recommendation system cannot be neglected any more. Current solutions to massive data processing problems include parallel processing (e.g. the distributed computing platform Hadoop) [43], fast clustering algorithm [55], Threshold-based Algorithm (TA in short) and improved TA algorithms [16], [44] (a query processing technology), heuristic methods [30] and so on.

Mahout [94] is an open source project belonging to the Apache Software Foundation (ASF) that includes many implementations such as clustering, classification, recommendation filtering, frequent itemset mining, etc. It relies on the current popular big data platform Hadoop and uses parallel distributed computing methods to deal with massive news data. If the news recommendation system is deployed on the Mahout platform, then the data processing speed requirement of the system can be satisfied [43]. At present, many algorithms in Mahout have been used to analyze user information, so as to realize the highly efficient personalized news recommendation. Among them, the slope-one algorithm is a lightweight CF recommendation algorithm. It assumes that there is a certain linear relationship between the preference values of two items, that is, the preference value of item j can be estimated by the preference value of item i through a linear function. The advantage of the algorithm is that its performance will not be affected by the number of users. Its performance only depends on the average difference of preference values between items. The difference value can be pre-calculated, and the underlying data structure can update efficiently. Therefore, the calculation is simple and fast, the update speed and scalability is good, and it is very suitable to use in the actual project. The disadvantage of the slope-one algorithm is that it is only applicable to the case where users' rating values on items are numerical, so the application range is limited.

In order to cope with massive news data and improve data processing speed, researchers have been trying to apply clustering methods to news recommendation techniques and have achieved some progress, such as K-means, fuzzy K-means, Locality Sensitive Hashing (LSH), the graph summarization algorithm SNAP-Cluster, probability-based topic model, etc. Among them, K-means is the most widely used clustering algorithm in machine learning, which is simple and easy to implement. In personalized news recommendation, its clustering effect and speed can be improved by changing the way of center selection or distance measurement to enhance the scalability. But its disadvantage is that it cannot ensure the

recommendation diversity, and then fuzzy K-means is proposed to solve this problem. LSH is a clustering algorithm by using hash operations to achieve reduction of feature dimension and fast clustering. The dimensionality reduction process can reduce the amount of computation greatly without affecting the original similarity between news texts. SNAP-Cluster is a clustering method which clusters user nodes based on the k-SNAP (k- Summarization by Grouping Nodes on Attributes and Pairwise Relationships) [66] graph summarization algorithm. It summarizes the user interests graph by k-SNAP to implement the interests clustering of different user nodes, and each node cluster is a series of aggregation of nodes with similar degrees of connectivity to other nodes [36]. This algorithm guarantees a certain recommendation diversity as well as fast clustering, but the iterative process of the algorithm is complicated somewhat. The typical representative of the probability-based topic models is LDA (Latent Dirichlet Allocation), each class of which is called a hidden class, and each text has probabilistic distributions on different hidden classes that express the probability of them belonging to the hidden class. There are many applications using this method in news recommendation. Some of the clustering algorithms listed above can be used directly in Mahout, while others still need to be integrated into recommendation algorithms. At present, in addition to the above clustering methods, density peaks clustering [65] and dynamic threshold network clustering [38] methods can also apply to the news data processing.

In addition to the use of parallel computing and fast clustering methods, Yin *et al.* [16], [44] propose to use a query processing technique to support the fast recommendation of massive news data online, and use the improved TA algorithm to enhance the processing speed of online recommendation, which can increase the recommendation efficiency. In order to make MobiFeed [30] adapt to the increasing amount of data in the system, a heuristic news feed method is designed in [30]. That is, the approximate optimal solution is found by intuition or empiricism at first, which is able to reduce the amount of computation effectively in the case of the data size becoming larger and larger. In [39], researchers exploit the reinforcement learning idea to transform the problems that are difficult for the original methods into the easy-to-handle problems. The behavior information are embedded in the generated continuous space, thus the complexity of processing behavior sets and the time complexity are reduced. At present, it is still an important task to study the scalable personalized news recommendation technology that can handle the huge amount of data.

7) TECHNOLOGIES FOR PRIVACY PROBLEM

Up to now, a large amount of work has been done to point out that recommender systems will cause privacy leaks as well as providing users with recommendation services. That is, the third-party servers or attackers can obtain user preference information in the recommendation process [68], [69]. At the same time, it is pointed out that the key issue of designing a recommender system lies in how to achieve a balance among the privacy, the utility and the cost in [70]. Therefore, the privacy protection in personalized recommendation process has become the current research hotspot, and researching on privacy protection schemes with good performance is an effective way to solve this problem. In [71], [72], a privacy-preserving mechanism based on anonymity scheme for context-aware recommender systems in cloud computing environment is proposed. Guo *et al.* [73] propose a trust-based fine-grained privacy protection mechanism in social network recommendation. In [74], users are classified into two kinds according to their privacy requirements. For a few ''public'' users, measures for protecting their privacy will not be taken in order to ensure the high recommendation accuracy; while for the most ''private'' users, the recommender system will take appropriate measures to protect their privacy by sacrificing a certain degree of accuracy. Researchers in [75], [76] suggest to design a formal privacy protection mechanism based on the differential privacy method. And, some cryptography-based schemes are also proposed to protect the privacy information of users in the recommender system in [77], [78]. Generally speaking, the homomorphic encryption method can be used for the protection of data in the system, and secure multi-party communication protocols can be adopted for the protection of the recommendation phase. In addition, there are many cryptographic techniques that can be used to alleviate the recommendation privacy issues. Besides, Zhu *et al.* [79] have designed a personalized privacy-aware App recommendation system, which is based on the popularity of the mobile terminal Apps and the privacy requirements of users. However, these schemes all have the problems of low recommendation quality and long response time, as they will introduce noise or use the homomorphic encryption technology.

Like the personalized recommendation technology in other areas, in the personalized news recommendation field, whether it is based on the users' historical behaviors or based on the news contents to analyze the correlations between users and news, it will always involve the same issue—the privacy issue. Considering this, Karkali *et al.* [40] point out that the recommender system should explicitly request permissions from users when collecting their sensitive data, and should process the data collected at the local client side, and not store the user data centrally to respect users' privacy. In addition, two recommendation methods are proposed in [62] without revealing the sensitive connections between users, and the methods make a compromise between recommendation accuracy and privacy protection. However, there is no more specific way to solve the privacy problem in news field at present, so it is a long-term and nonnegligible research direction of all recommender systems.

C. DISPLAY OF PERSONALIZED NEWS RECOMMENDATION RESULTS

How to present news recommendation results to users effectively and reasonably to provide a better user experience is

also an important step in recommender system. Two display ways of news in personalized news recommendation are given bellow separately considering that the users may read news on the mobile terminal or the web side.

1) MOBILE SIDE

Due to the limited screen size of mobile devices, simplicity is required firstly in the design of the news recommendation results presentation, followed by adjustable parameters, interoperability, exploratoryness and so on. And there is a requirement that users are able to directly modify parameters such as time, place, news source, news topics, news categories, quick updates of related news, etc. when they are browsing news pages. The ability to get the entire news quickly from the brief news displayed on the mobile side is also required.

The limitations of display in mobile devices can be overcome by the following methods. The first method is that news interfaces with different granularity are divided to display according to the amount of news information that interfaces contain, such as the coarse-grained news overall interface and the fine-grained single news display interface, etc. The second method is to replace UI designs which occupy large interface space like buttons with user actions such as sliding and double-clicking, etc. The third method is to display only specific contents at specific time and specific scenes to achieve the efficient and reasonable utilization of time slices.

The pull-to-refresh and sliding-side-menu approaches proposed in [19] can improve users' satisfaction. Considering the display contents, it selects news titles, summaries and background colors representing different news categories to display. Considering the way of display, it chooses two types: plate and map ways. In addition, the titles, short descriptions, URLs and publication dates of the recommendation news are selected to display in the list way in [40]. The list way is similar to the plate way, they are both flexible to use, while the map way is more suitable for news recommendation based on the location context so that the news and the geographical location can be linked directly and intuitively. Moreover, personalized recommendation for news radio has also been proposed in [10]. As a display way, it is able to meet the needs of users who cannot or inconveniently use the screen to read news. Compared with the previous screen display way, there are less or even no need for users to interact with the system at all, so this approach can play a very good role in specific fields.

2) WEB SIDE

Web-side news display has no restrictions as the mobile side. The news matching with user interests and the latest, hottest news can be selected flexibly to display on the web side. A better approach is to display according to categories, such as sports, entertainment, political commentary, financial information, cutting-edge science, art and so on, which not only meets the interest relevance, but also meets the diversity, bringing users a good use experience. The well-known domestic news websites are NetEase, SOHU, today's headline, etc. MicroBlog belongs to the social media category, but it also has the functions of news feeds and news propagation. The well-known news websites abroad are Google News, Yahoo! News, BBC, New York Times and so on. Different news websites have different characteristics and focus on different aspects, but the display ways are generally similar, as the detailed news pages all have a large number of pictures, some text fragments, even dynamic images and short videos.

VI. OPEN ISSUES IN PERSONALIZED NEWS RECOMMENDATION

With the development of text mining and information retrieval technologies over time, the content-based recommendation method already has a considerable advantage for the recommendation of textual information, which promotes the continuous development and improvement of personalized recommendation techniques in the field of news to some extent. However, there are still many challenges in the research of personalized news recommendation technology.

A. DATA SPARSENESS AND COLD START PROBLEMS

Data sparseness and cold start problems are the unavoidable problems for all recommender systems. In the field of personalized news recommendation, users' behavior data like reading, collecting and commenting on massive Internet news are very limited, making the user-news rating matrix extremely sparse, and resulting in a low accuracy of the similarity calculation between the users and the news, and finally causing a low recommendation performance. In the meantime, new users and new news cannot be recommended effectively because they lack the necessary historical behavior data to build models for them.

B. SCALABILITY AND EFFICIENCY ISSUES

With the increasing number of Internet users and the drastic increase of the amount of news data, the pressure of data processing in personalized news recommendation algorithm is increasing too. And the scalability of the system is highlighted. At present, researchers mainly relieve the scalability problem from several aspects, including reducing datasets, fast clustering, degrading the dimension of features and establishing a linear model, etc. However, with the introduction of users' social network information and contextual information, the scale of dataset will be further expanded. This phenomenon will inevitably further increase the data pressure and computational complexity of the personalized news recommendation algorithm. And it may be the bottleneck restricting the overall performance and efficiency of personalized news recommendation. Therefore, how to design the fast recommendation algorithm that requires high real-time effect and high efficiency is a key point in the research of personalized news recommendation technology at present.

C. PERSONALIZED NEWS RECOMMENDATION INTEGRATED WITH OTHER INFORMATION

With the rapid development of social networks and mobile networks, the breadth and depth of Internet news dissemination have reached an unprecedented scale. As a result, there have been technologies that integrate social network

information and mobile contextual information into personalized news recommendation, which are respectively referred to as social network news recommendation technology and context-based news recommendation technology. Users in social network are more active, and have frequent interactions with other users, and it is easy to form popular information in a short period of time [61]. Therefore, considering the impact of the hot topics appearing in the social network on the short-term interest of users is of great help for the news recommendation. At the same time, personalized news recommendation technologies are also affected by different types of contextual information to a different extent, especially in the mobile news recommendation. Considering the time, location and other contextual factors of users reading news is of great significance in mining user interests and constructing user models, but due to the variability of the environment, it is very difficult to make context-aware recommendation. Now the two directions above are the research hot spots.

D. PRIVACY PROTECTION AND SECURITY ISSUES

Making personalized news recommendation needs to collect a large number of user personal information, including the users' reading history records, demographic information, social relations, location information, etc. Once the information are leaked or maliciously stolen, it will bring great security risks, threatening users' property safety or even their life safety, so the privacy protection [40], [62] problem in the news recommendation cannot be underestimated. In addition, the personalized news recommendation also has the security problem of recommending attacks. That is, attackers maliciously create fake rating data by some methods like injecting fake users into the system. Users' cheating behaviors on rating data have always been an important issue restricting the development of recommendation technologies. Some people [63] propose a unified framework based on the proliferation of suspicious behaviors to make up for the deficiency that the traditional anti-cheating methods need to design different anti-cheating algorithms for different cheating behaviors. The framework allows the system maintenance personnel to identify the cheating users and their cheating behaviors with a relatively high accuracy, but without concerning for the specific cheating methods, to remove the adverse effects.

E. INFORMATION COCOONS

The concept of ''Information Cocoons'' [64] was first proposed by Sunstein, a professor at the University of Chicago in the United States. ''In the information dissemination, as the information needs of the public are not comprehensive, that is, the public only pays attention to what they choose and the communication field making them pleasurable, gradually, they will be lingering in their own 'cocoons', like a cocoon room.'' ''Information Cocoons'' is also called ''Information Filtering Bubbles''. The contents of some news websites on the Internet are relatively simple, which easily leads to the one-sided and inaccurate user interest models, and may even leads to failing to dig out the potential interests of users, and finally affecting the diversity and novelty of recommendation results seriously. Moreover, completely personalized news information is equivalent to cutting off the personal exposure opportunities to the diversified information all over the world. The concentration of a large number of similar information will greatly narrow users' perspective, which is not conducive to the full development of people and may even results in an imbalance on the personal information structure. Recently, researchers have realized the drawbacks of over-personality and are taking measures aggressively. For example, the operators of today's headline reinforce the concept of model generalization [95] in their recommendation algorithms, trying to free users from the information cocoons, but the effect is not very good.

F. THE PROBLEM OF MISSING VALUE JUDGMENTS

Traditional news needs to be filtered by editors before being published, and only the news that meets the requirements can be released. In this situation, the editors act as a ''gatekeeper'' in order to filter out some bad information for readers. However, the ways of information filtering in personalized news recommendation algorithms are mainly to avoid certain keywords and rely on a huge database foundation, which cannot guarantee the quality of the news, and many undesirable news information will still enter users' field of vision. Therefore, the existing recommendation algorithms still have discrepancies in filtering news, as it is difficult to judge the value and discriminate between true and false in news, so the algorithms in personalized news recommendation still need further studies.

VII. CONCLUSION

In the 21st century, the information on the Internet develops rapidly, and the Internet users are experiencing an ever-increasing demand for information. As a result, the problem of information overload is getting more and more serious. The role of personalized recommendation in the digitalization process of all fields on the Internet is also becoming increasingly important, especially, in the news field. It is of great benefit for news readers, news writers and news feed platforms to do such work. One is to integrate the existing personalized recommendation technologies into Internet news; the second is to build the optimal user preference model considering multiple factors comprehensively; and the third is to find the data mining methods that are most suitable for news data characteristics. What's more, the work can mitigate the negative impact on the news area caused by information overload problem, such as the serious user churn problem. This paper analyzes the similarities and differences between personalized news recommendation techniques and personalized recommendation techniques in other fields first, and then mainly introduces and analyzes the research status and progress of the news recommendation techniques for different characteristics of news, and finally points out some open issues, hoping that it can provide some useful help to researchers and enterprise engineering technical personnel in related fields of news recommendation.

ACKNOWLEDGMENT

The authors would like to thank anonymous reviewers for their valuable comments.

REFERENCES

[1] L. Xiang, *Recommender System in Practice*. Beijing, China: Posts & Telecom Press (in Chinese), 2012.
[2] J. Liu, *The Principle of Contemporary Journalism*. Beijing, China: Tsinghua Univ. Press, (in Chinese), 2003.
[3] N. Manouselis and C. Costopoulou, "Analysis and classification of multi-criteria recommender systems," *World Wide Web*, vol. 10, no. 4, pp. 415–441, 2007.
[4] K. F. Yeung, Y. Yang, and D. Ndzi, "A proactive personalised mobile recommendation system using analytic hierarchy process and Bayesian network," *J. Internet Services Appl.*, vol. 3, no. 2, pp. 195–214, 2012.
[5] F. Bodon, "A fast APRIORI implementation," in *Proc. IEEE ICDM Workshop Frequent Itemset Mining Implement.*, Melbourne, FL, USA, Nov. 2003, pp. 14–23.
[6] Y. Niu, Y. Wang, and G. Sun, "The Tencent dataset and KDD-Cup'12," in *Proc. KDD-Cup Workshop*, Beijing, China, 2012, pp. 1–6.
[7] S.-Y. Song and Q.-D. Li, "Micro-blogging information recommendation system for mobile client," (in Chinese), *Comput. Sci.*, vol. 38, no. 11, pp. 137–141, 2011.
[8] C.-M. Chen, C.-M. Hong, and S.-C. Chen, "Intelligent location-based mobile news service system with automatic news summarization," in *Proc. Int. Conf. Environ. Sci. Inf. Appl. Technol.*, Jul. 2009, pp. 527–530.
[9] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proc. 19th Int. Conf. World Wide Web*, Raleigh, NC, USA, 2010, pp. 661–670.
[10] Y. Xie, L. Chen, K. Jia, L. Ji, and J. Wu, "iNewsBox: Modeling and exploiting implicit feedback for building personalized news radio," in *Proc. ACM CIKM*, 2013, pp. 2485–2488.
[11] W.-S. Yang, H.-C. Cheng, and J.-B. Dia, "A location-aware recommender system for mobile shopping environments," *Expert Syst. Appl.*, vol. 34, no. 1, pp. 437–445, 2008.
[12] J. L. Hong, J. Y. Choi, and S. J. Park, "Context-aware recommendations on the mobile Web," in *Proc. OTM Confederated Int. Workshops Posters*, Agia Napa, Cyprus, 2005, pp. 142–151.
[13] L. Wang and X. Meng, "A MAUT approach to elicitation of contextual user preferences," *Adv. Inf. Sci. Service Sci.*, vol. 4, no. 5, pp. 98–105, 2012.
[14] I. Cantador, A. Bellogín, and P. Castells, "A multilayer ontology-based hybrid recommendation model," *AI Commun.*, vol. 21, pp. 203–210, Jan. 2008.
[15] Y. Wu, Y. Ding, X. Wang, and J. Xu, "Topic based automatic news recommendation using topic model and affinity propagation," in *Proc. 9th Int. Conf. Mach. Learn. Cybern.*, Qingdao, China, 2010, pp. 11–14.
[16] H. Yin, Y. Sun, B. Cui, Z. Hu, and L. Chen, "LCARS: A location-content-aware recommender system," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 221–229.
[17] J. W. Son, A. Y. Kim, and S. B. Park, "A location-based news article recommendation with explicit localized semantic analysis," in *Proc. ACM SIGIR*, 2013, pp. 293–302.
[18] P.-H. Chiu, G. Y.-M. Kao, and C.-C. Lo, "Personalized blog content recommender system for mobile phone users," *Int. J. Hum.-Comput. Stud.*, vol. 68, no. 8, pp. 496–507, 2010.
[19] M. Tavakolifard, J. A. Gulla, and K. C. Almeroth, "Tailored news in the palm of your HAND: A multi-perspective transparent approach to news recommendation," in *Proc. WWW Companion*, 2013, pp. 305–308.
[20] Y. Wu, M. Qi, and R. Yang, "A news recommendation system based on an improved collaborative filtering algorithm," (in Chinese), *Comput. Eng. Sci.*, vol. 39, no. 6, pp. 1179–1185, 2017.
[21] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, Aug. 2009.
[22] H. Ma, I. King, and M. R. Lyu, "Learning to recommend with social trust ensemble," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2009, pp. 203–210.
[23] X. Song, B. L. Tseng, C. Y. Lin, and M.-T. Sun, "Personalized recommendation driven by information flow," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Seattle, WA, USA, Aug. 2006, pp. 509–516.
[24] Y. Seroussi, F. Bohnert, and I. Zukerman, "Personalised rating prediction for new users using latent factor models," in *Proc. HT*, 2011, pp. 47–56.
[25] C. Lin, R. Xie, L. Li, Z. Huang, and T. Li, "PRemiSE: Personalized news recommendation via implicit social experts," in *Proc. Int. Conf. Inf. Knowl. Manage.*, 2012, pp. 1607–1611.
[26] L. Baltrunas, B. Ludwig, S. Peer, and F. Ricci, "Context relevance assessment and exploitation in mobile recommender systems," *Pers. Ubiquitous Comput.*, vol. 16, no. 5, pp. 507–526, 2012.
[27] L. Li, L. Zheng, F. Yang, and T. Li, "Modeling and broadening temporal user interest in personalized news recommendation," *Expert Syst. Appl.*, vol. 41, no. 7, pp. 3168–3177, 2014.
[28] F. Garcin, C. Dimitrakakis, and B. Faltings, "Personalized news recommendation with context trees," in *Proc. RecSys*, 2013, pp. 105–112.
[29] J. Bao, M. F. Mokbel, and C.-Y. Chow, "GeoFeed: A location-aware news feed system," in *Proc. IEEE Int. Conf. Data Eng.*, Apr. 2012, pp. 54–65.
[30] W. Xu, C.-Y. Chow, M. L. Yiu, Q. Li, and C. K. Poon, "MobiFeed: A location-aware news feed system for mobile users," in *Proc. Int. Conf. Adv. Geograph. Inf. Syst.*, 2012, pp. 538–541.
[31] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proc. Int. Conf. World Wide Web*, 2010, pp. 591–600.
[32] K. Lerman and R. Ghosh, "Information contagion: An empirical study of the spread of news on digg and Twitter social networks," *Comput. Sci.*, vol. 52, pp. 166–176, Mar. 2012.
[33] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: Real-time event detection by social sensors," in *Proc. Int. Conf. World Wide Web*, 2010, pp. 851–860.
[34] O. Phelan, K. Mccarthy, and B. Smyth, "Using Twitter to recommend real-time topical news," in *Proc. ACM Conf. Recommender Syst.*, 2009, pp. 385–388.
[35] W.-J. Lee, K.-J. Oh, C.-G. Lim, and H.-J. Choi, "User profile extraction from Twitter for personalized news recommendation," in *Proc. Int. Conf. Adv. Commun. Technol.*, 2014, pp. 779–783.
[36] K.-H. Chen, P.-P. Han, and J. Wu, "Heterogeneous social network recommendation algorithm based on user clustering," (in Chinese), *Chin. J. Comput.*, vol. 36, no. 2, pp. 349–359, 2013.
[37] X. Chen, P. Wang, Z. Qin, and Y. Zhang, "HLBPR: A hybrid local Bayesian personal ranking method," in *Proc. Int. Conf. Companion World Wide Web*, 2016, pp. 21–22.
[38] L. Zhang, Y. Hu, and S. Lu, "Exploring the core courses and the course structures for undergraduate education," *Chin. Sci. Bull.*, vol. 62, nos. 28–29, pp. 3277–3284, 2017.
[39] G. Dulac-Arnold, R. Evans, H. van Hasselt, P. Sunehag, T. Lillicrap, J. Hunt, T. Mann, T. Weber, T. Degris, and B. Coppin, "Deep reinforcement learning in large discrete action spaces," 2016, *arXiv:1512.07679*. [Online]. Available: https://arxiv.org/abs/1512.07679
[40] M. Karkali, D. Pontikis, and M. Vazirgiannis, "Match the news: A firefox extension for real-time news recommendation," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2013, pp. 1117–1118.
[41] L. Li, W. Chu, J. Langford, and X. Wang, "Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms," in *Proc. ACM Int. Conf. Web Search Data Mining*, 2011, pp. 297–306.
[42] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in *Proc. Conf. Uncertainty Artif. Intell.*, 2009, pp. 452–461.
[43] J. Yin, Z.-S. Wang, Q. Li, and W.-J. Su, "Personalized recommendation based on large-scale implicit feedback," (in Chinese), *J. Softw.*, vol. 25, no. 9, pp. 1953–1966, 2014.
[44] H. Yin, B. Cui, L. Chen, Z. Hu, and Z. Huang, "A temporal context-aware model for user behavior modeling in social media systems," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2014, pp. 1543–1554.
[45] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," in *Multimedia Services in Intelligent Environments*. Springer, 2013, pp. 734–749.
[46] L. Li, D. Wang, T. Li, D. Knox, and B. Padmanabhan, "SCENE: A scalable two-stage personalized news recommendation system," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Beijing, China, 2011, pp. 125–134.
[47] B. M. Sarwar, G. Karypis, and J. Konstan, "Analysis of recommendation algorithms for e-commerce," in *Proc. 2nd ACM Conf. Electron. Commerce*, 2000, pp. 158–167.
[48] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: An open architecture for collaborative filtering of netnews," in *Proc. ACM Conf. Comput. Supported Cooperat. Work (CSCW)*, 1994, pp. 175–186.

[49] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proc. 10th Int. Conf. World Wide Web*, 2001, pp. 285–295.
[50] A. Gunawardana and G. Shani, "A survey of accuracy evaluation metrics of recommendation tasks," *J. Mach. Learn.*, vol. 10, pp. 2935–2962, Dec. 2009.
[51] Y. Zhang, Y. Zhuang, J. Wu, and L. Zhang, "Applying probabilistic latent semantic analysis to multi-criteria recommender system," *AI Commun.*, vol. 22, no. 2, pp. 97–107, 2009.
[52] H. Kautz, B. Selman, and M. Shah, "Referral Web: Combining social networks and collaborative filtering," *Commun. ACM*, vol. 40, no. 3, pp. 63–65, 1997.
[53] P. Melville, R. J. Mooney, and R. Nagarajan, "Content-boosted collaborative filtering for improved recommendations," in *Proc. 18th Nat. Conf. Artif. Intell.*, 2002, pp. 187–192.
[54] R. Burke, "Hybrid recommender systems: Survey and experiments," *User Model. User-Adapt. Interact.*, vol. 12, no. 4, pp. 331–370, Nov. 2002.
[55] M. O'Cornor and J. Herlocker. "Clustering items for collaborative filtering," in *Proc. ACM SIGIR Workshop*, 1999, pp. 1–4.
[56] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, "Methods and metrics for cold-start recommendations," in *Proc. 25th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2002, pp. 253–260.
[57] F. Abel, Q. Gao, G.-J. Houben, and K. Tao, "Analyzing user modeling on Twitter for personalized news recommendations," in *Proc. 19th Int. User Modeling, Adaption Personalization (UMAP) Conf.*, Girona, Spain, Jul. 2011, pp. 1–12.
[58] M. Capelle, F. Frasincar, M. Moerland, and F. Hogenboom, "Semantics-based news recommendation," in *Proc. 2nd Int. Conf. Web Intell., Mining Semantics*, 2012, p. 27.
[59] H. Ren and W. Feng, "CONCERT: A concept-centric Web news recommendation system," in *Web-Age Information Management*. Berlin, Germany: Springer, 2013, pp. 796–798.
[60] N. Moniz, L. Torgo, and M. Eirinaki, "Socially driven news recommendation," 2016, *arXiv:1506.01743*. [Online]. Available: https://arxiv.org/abs/1506.01743
[61] L. Quijano-Sanchez, J. A. Recio-Garcia, B. Diaz-Agudo, and G. Jimenez-Diaz, "Social factors in group recommender systems," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 1, 2013, Art. no. 8.
[62] A. Machanavajjhala, A. Korolova, and A. Sarma, "Personalized social recommendations: Accurate or private?" *Proc. VLDB Endowment*, vol. 4, no. 7, pp. 440–450, 2011.
[63] Y. Zhang, Y. Tan, M. Zhang, Y. Liu, T.-S. Chua, and S. Ma, "Catch the black sheep: Unified framework for shilling attack detection based on fraudulent action propagation," in *Proc. Int. Conf. Artif. Intell.*, 2015, pp. 2408–2414.
[64] C. R. Sunstein, *Infotopia*. Beijing, China: China Law Press, 2008.
[65] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014.
[66] Y. Tian, R. A. Hankins, and J. M. Patel, "Efficient aggregation for graph summarization," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Vancouver, Bc, Canada, Jun. 2008, pp. 567–580.
[67] H.-S. Zhang, G.-S. Chen, Y.-T. Ma, and Y.-C. Liu, "Group interests and their correlations mining based on Wikipedia," (in Chinese), *Chin. J. Comput.*, vol. 34, no. 11, pp. 2234–2242, 2011.
[68] J. A. Calandrino, A. Kilzer, A. Narayanan, and E. W. Felten, "'You might also like:' Privacy risks of collaborative filtering," in *Proc. IEEE Symp. Secur. Privacy (S&P)*, Berkeley, CA, USA, May 2011, pp. 231–246.
[69] S. Bhagat, U. Weinsberg, S. Ioannidis, and N. Taft, "Recommending with an agenda: Active learning of private attributes using matrix factorization," in *Proc. 8th ACM Conf. Recommender Syst.*, Foster City, CA, USA, 2014, pp. 65–72.
[70] C. Staff, "Recommendation algorithms, online privacy, and more," *Commun. ACM*, vol. 52, no. 5, pp. 10–11, 2009.
[71] A. H. Celdrán, M. G. Pérez, F. J. G. Clemente, and G. M. Pérez, "PRECISE: Privacy-aware recommender based on context information for cloud service environments," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 90–96, Aug. 2014.
[72] J. Zhu, P. He, Z. Zheng, and M. R. Lyu, "A privacy-preserving QoS prediction framework for Web service recommendation," in *Proc. IEEE Int. Conf. Web Services*, Jun./Jul. 2015, pp. 241–248.
[73] L. Guo, C. Zhang, and Y. Fang, "A trust-based privacy-preserving friend recommendation scheme for online social networks," *IEEE Trans. Dependable Secure Comput.*, vol. 12, no. 4, pp. 413–427, Jul. 2015.
[74] X. Yu and T. S. Jaakkola, "Controlling privacy in recommender systems," in *Proc. Int. Conf. Neural Inf. Process. Syst.* Cambridge, MA, USA: MIT Press, 2014, pp. 2618–2626.
[75] R. Guerraoui, A.-M. Kermarrec, R. Patra, and M. Taziki, "D2P: Distance-based differential privacy in recommenders," *Proc. VLDB Endowment*, vol. 8, no. 8, pp. 862–873, 2015.
[76] Y. Shen and H. Jin, "Privacy-preserving personalized recommendation: An instance-based approach via differential privacy," in *Proc. IEEE Int. Conf. Data Mining*, Shenzhen, China, Dec. 2014, pp. 540–549.
[77] B. K. Samanthula, L. Cen, W. Jiang, and L. Si, "Privacy-preserving and efficient friend recommendation in online social networks," *Trans. Data Privacy*, vol. 8, no. 2, pp. 141–171, 2015.
[78] T. R. Hoens, M. Blanton, A. Steele, and N. V. Chawla, "Reliable medical recommendation systems with patient privacy," in *Proc. ACM Int. Health Inform. Symp.*, 2010, pp. 173–182.
[79] H. S. Zhu, H. Xiong, Y. Ge, and E. Chen, "Mobile app recommendations with security and privacy awareness," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 951–960.
[80] *InfoChimps Datasets*. Accessed: Jan. 2018. [Online]. Available: http://www.infochimps.com/
[81] *Kaggle Datasets*. Accessed: Jan. 2018. [Online]. Available: https://www.kaggle.com/datasets
[82] *Million Song Dataset*. Accessed: Jan. 2018. [Online]. Available: https://labrosa.ee.columbia.edu/millionsong/lastfm
[83] *Book-Crossing Dataset*. Accessed: Jan. 2018. [Online]. Available: http://www2.informatik.uni-freiburg.de/~cziegler/BX/
[84] *MovieLens Dataset*. Accessed: Jan. 2018. [Online]. Available: https://grouplens.org/datasets/movielens/
[85] *Recommendation Website*. Accessed: Jan. 2018. [Online]. Available: https://www.outbrain.com/
[86] *Frequent Itemset Mining Dataset Repository*. Accessed: Jan. 2018. [Online]. Available: http://fimi.ua.ac.be/data/
[87] *Wikipedia Website*. Accessed: Jan. 2018. [Online]. Available: https://www.wikipedia.org/
[88] *Newsgroup Binaries*. Accessed: Jan. 2018. [Online]. Available: http://www.newsgroup-binaries.com/
[89] *Junar Data Platform*. Accessed: Jan. 2018. [Online]. Available: http://www.junar.com/index9ed2.html?lang=en
[90] *Datasets for Investigating Recommender Systems*. Accessed: Jan. 5, 2018. [Online]. Available: https://gab41.lab41.org/the-nine-must-have-datasets-for-investigating-recommender-systems-ce9421bf981c
[91] *Anonymous Ratings From the Jester Online Joke Recommender System*. Accessed: Jan. 2018. [Online]. Available: http://eigentaste.berkeley.edu/dataset/
[92] *Baike Baidu*. Accessed: Jan. 5, 2018. [Online]. Available: https://baike.baidu.com/item/%E6%96%B0%E9%97%BB/289002?Fr=aladdin
[93] *SimilarWeb*. Accessed: Jan. 23, 2018. [Online]. Available: http://group.jobbole.com/10838/
[94] *Mahout Official Website*. Accessed: Jan. 2018. [Online]. Available: http://mahout.apache.org/
[95] *Douban*. Accessed: Jan. 7, 2018. [Online]. Available: https://site.douban.com/250906/widget/notes/18386315/note/619519047/

MIAOMIAO LI received the B.S. degree from the North China University of Water Resources and Electric Power, in 2016. She is currently pursuing the M.S. degree with the Beijing University of Posts and Telecommunications. Her current research interests include personalized recommendation technology, the security of data mining algorithm, and network security.

LICHENG WANG received the B.S. degree from Northwest Normal University, in 1995, the M.S. degree from Nanjing University, in 2001, and the Ph.D. degree from Shanghai Jiao Tong University, in 2007. He is currently an Associate Professor with the Beijing University of Posts and Telecommunications. His current research interests include modern cryptography, network security, and trust management.