Research Article

A Computational Complexity-Based Method for Predicting Scholars’ Ages through Articles’ Information

Jun Zhang,1 Xiaoyan Su,2 Mingliang Hou,3 and Jing Ren3

1Graduate School of Education, Dalian University of Technology, Dalian 116024, China
2School of Economics and Management, Dalian University of Technology, Dalian 116024, China
3School of Software, Dalian University of Technology, Dalian 116620, China

Correspondence should be addressed to Jun Zhang; junzhang@dlut.edu.cn

Received 13 October 2020; Revised 29 December 2020; Accepted 1 April 2021; Published 20 April 2021

Copyright © 2021 Jun Zhang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Many scholars have conducted in-depth research on the evaluation and prediction of scholars’ scientific impact and meanwhile discovered various factors that affect the success of scholars. Among all these relevant factors, scholars’ ages have been universally acknowledged as one of the most important factors for it can shed light on many practical issues, e.g., finding supervisors, discovering rising stars, and research funding or award applications. However, due to the inaccessibility or the privacy issues of acquiring scholars’ personal data, there is little research to explore the true ages of scholars currently. Alternatively, scholars’ publications’ information can be obtained through various digital libraries. Inspired by this fact, we propose a novel scholar’s age prediction method based on their articles’ information. Our method first classifies factors that affect scholars’ ages into intuitive and complex types according to their computational complexity and then apply machine learning algorithms to predict the ages of scholars based on these factors. The experimental results on the real dataset demonstrate that our method can effectively predict the true ages of scholars. Given that there is no completely accurate dataset because of the continuous publication of academic papers, we then apply our method on the incomplete dataset. Nevertheless, our method still has high prediction accuracy in such situations.

1. Introduction

With the rapid growth of scholarly big data, evaluating and predicting the scientific impact are of great significance since it can shed light on many practical issues, such as ranking institutions, providing basis for research funding’s assignment, hiring faculties, and furthermore, promoting the advancement of science [1, 2]. As a consequence, more and more scholars have begun to explore the factors that affect the success of scholars [3]. Among them, many studies have confirmed the crucial role of scholars’ actual ages on their future achievements, and it can also provide basis for making academic decisions [4, 5].

Recognizing the age of scholars, a series of academic questions can be answered. For instance, it is important for students to find appropriate supervisors or mentors who will play vital roles during their whole academic career in many cases, such as fostering students’ specialization interests, guiding their research directions, and even providing some helpful advice for their job selections. Knowing supervisors’ ages in advance can offer basis for students when making their own academic plan at school. Meanwhile, scholars’ ages are also an important index when apply for research grants or academic awards. Besides, the fundamental component of a research institution or university is scholars. Their research and teaching capacities can affect the development of institutions to a large extent. Thus, many universities have tried diverse ways to optimize their teachers’ age structures including searching and recruiting young academic rising stars, and knowing scholars’ ages can provide basis for such questions. Moreover, from the aspect of economics, knowing the age of a scientist gives us important information about state of matters about scientific education, costs of knowledge accumulation, and changing productivity.
patterns over the life cycle of scientists, especially because strong cohort effects are at play due to the increasing burden of knowledge [6].

However, due to some privacy issues, many scholars do not disclose their true ages, and consequently discovering the true ages of scholars is one of the crucial problems to be solved. Based on the abovementioned facts, the main content of this paper is to predict scholars’ true ages according to their publications’ information.

Most of the current research work on predicting the scholars’ true ages originates from the study of user attribute mining in social networks [7]. With the growing popularity of social media, more and more people are joining and using social media networks to share their daily lives and browse the news they are interested in. Facebook now has nearly 2 billion users who continually create and share large-scale of multimedia information, such as posts, photos, and videos. YouTube uploads videos about 100 hours per minute and more than 6 billion hours per month, and similarly, Facebook users have uploaded more than 250 billion photos. Such rich multimedia information can provide vast important clues for exposing the users’ information, including age, gender, personal interests, occupations, and others. Analysis of these users’ attribute information can contribute to many services such as commodity recommendations of shopping websites [8]. However, most online social networks, such as Facebook, Twitter, and SinaWeibo, do not directly offer users’ attribute information in order to protect their privacy. As a consequence, researchers began to exploit data from social networks to seek out user attributes [9, 10]. Based on a large amount of data on social networks, researchers try to explore user attributes from different perspectives and analyze their behaviors through these attributes, for instance, exploring the diverse language usage habits among Facebook users by genders and organizations. Researchers also predict some personal attributes, such as races, religions, political opinions, and even sexual orientations based on social media records [11, 12]. In addition to mining user attributes, researchers can analyze and predict user behaviors through data on social networks at the same time [13, 14].

One of the important sociodemographic characteristics is people’s age, and it has been proven to be a key influence factor of research productivity [15]. However, due to the privacy of such data, obtaining the true ages of scholars directly is difficult, which leads to the lack of related research work. To compensate for this shortcoming, researchers use alternative methods to consider the influence of scholars’ ages. For instance, consider the academic ages of scholars and the time when scholars publish their first papers to study their impact on scientific influence or academic cooperation models [16]. Although these alternative factors have a certain correlation with age, they cannot reflect the true ages of scholars accurately. Meanwhile, unlike the user attribute mining problem in social networks where the amount and variety of data in social networks are very rich and closely related to users themselves, most of the scholar-related data that can be acquired are the information of their papers. In addition, scholars’ ages are significantly important in our country. For example, many universities have specific rules for scholars’ ages when recruiting faculties, e.g., the age of lecturer should be under 35 years old. Moreover, our country selects “first-class universities and disciplines” every five years. For those universities and disciplines that are selected in this project, they can get more financial and policy supports from the government. The age distribution of faculties within a university is an important indicator in the selecting process. Therefore, many universities have tried diverse ways to optimize their teachers’ age structures including searching and recruiting young academic rising stars. Furthermore, due to the continuous publication of academic articles, the incompleteness of a scholar’s whole publication’s information will increase the complexity of predicting scholars’ age [17].

Motivated by the abovementioned facts, the major work of this paper is predicting the actual ages of scholars through their papers’ information. Our proposed method consists of two main parts: discovering the relative factors and then utilizing the machine learning algorithms to predict scholars’ ages. Firstly, we utilize the publications’ data (title, abstract, authors, etc.) of scholars to explore factors correlating with the ages of scholars. For each factor, we propose a specific quantitative method to compute, and then the true ages of the scholars can be predicted by inputting these factors into machine learning algorithms while XGBoost, GBDT, and SVR algorithms have been widely used and recognized by the mainstream of the industry. Therefore, XGBoost, GBDT, and SVM algorithms are selected to do the prediction in this paper. In summary, we make the following contributions in this paper:

(i) We propose a novel scholar’s true age prediction method, which relies on scholars’ paper information rather than the multiple data types from social networks.
(ii) We conduct an in-depth exploration of the factors that affect the prediction of scholars’ ages and discover several closely related factors, including the publication time of the first paper and the distribution of collaborators.
(iii) Experiments performed on real datasets show that our proposed method can effectively predict the true ages of scholars. Furthermore, even though limited by the incompleteness of datasets, we can still achieve a high accuracy.

The rest of this paper is organized as follows: Section 2 presents the related work. Section 3 introduces our proposed method. Section 4 demonstrates the experimental results of our method. At last, we conclude our work in Section 5.

2. Related Work

With the surge of social networks, research on user attribute mining has attracted more and more attentions. Based on social network data, such as user postings or social relationships, researchers can predict user’s gender, religion, and age. The study of scholars’ age prediction problem stems
from the mining of user attributes in social networks [18, 19]. This section will present the current research work in the above areas.

2.1. User Attribute Mining. By utilizing the diverse social network data, researchers begin to predict the attributes of users based on the text information. Garera et al. [20] use text data to mine its structure and transfer properties to predict user attributes such as birthday, occupation, and nationality. Mann et al. [21] apply context learning methods to extract user attribute information, such as birthplace. Bergsma et al. [22] utilize the properties of the concept class to predict the hidden nature of social network users. Although the above work can effectively extract user attribute information using rules-based or pattern-based learning methods, learning patterns or rules are limited to specific attributes, which will result in their nonuniversality.

Therefore, some researchers try to combine the user communication information with online multimedia data to mine user attributes. User profiles greatly influence their characteristics and social networking activities [23]. These studies [24–26] have analyzed the relationship between user profiles and their social activities, and they provide a reliable basis for inferring their attributes through user data. Most of the current research work also solves such problems by designing user-related attributes and using existing classification algorithms. Garera et al. [27] extend the N-gram based model by leveraging sociolinguistic features and support vector machine models. Rao et al. [28] use the sociolinguistic features and the N-gram model to predict Twitter user genres. Bergsma et al. [22] propose a method based on communication data and location clustering from Twitter to improve the result of user classification. Based on the dynamic evolution of networks, Volkova et al. [29] propose a prediction method by applying Bayesian network incremental learning. Culotta et al. [30] use traffic data on the website to predict demographic characteristics.

When mining user attributes, current research focuses on presenting more complex feature-based classifiers. Rao et al. [28] propose a support vector machine (SVM) method based on social network user attributes. Rao et al. [31] propose a hierarchical Bayesian model for predicting the race and gender of users in Facebook, which applied the N-gram method to extract features from their Facebook information. By considering the closely related vocabularies to the region, geographically coherent linguistic regions, and the changes between regions and topics, Eisenstein et al. [32] propose a cascading topic model to classify users’ regions. Sun et al. [33] propose the Content Enhancement Network Embedding Method (CENE), which enables the collaborative use of social network structure information and content information in an unsupervised learning manner.

2.2. Age Mining. Based on the algorithmic foundation of the abovementioned attribute mining, researchers began to dig deeper into the age attribute. Researchers in [34, 35] observe that users’ interests can be explored by analyzing posts posted by their followers. It can be seen that on social network sites such as Twitter, Facebook, and YouTube, users’ interests are mainly influenced by their followers. As the volume of logs on networks increased, researchers also began to use the subnetwork of followers to predict the age of users [36]. In addition, the user’s age can also be predicted by analyzing the web browsing data through probabilistic machine learning algorithms. These methods, which can effectively perform on the large-scale data and obtain an accurate result, transform the age predicting work into predication problem or classification problem. The following research work will be introduced from the abovementioned two aspects.

Human ages can be identified from a series of images, for instance, each year’s image of a person is different from the previous. Inspired by this, Guo et al. [37] predict ages by designing a probability model to analyze multimedia data. In [38], features extracted from blogs and forums are taken into account, and the age of user is predicted by linear regression algorithm. Similar to [38], Nane et al. [39] also predict the scholars’ ages utilizing the linear regression algorithm from different disciplines.

Age prediction is also recognized as a classification problem. In [28], it predicts the age attribute of Twitter users based on the SVM classification method. The World Health Organization (WHO) provides a set of age groups that are dominant in the United States, but the ages of Twitter users based on age groups do not accurately reflect the user’s age. For example, in the 17 to 30 age group, the age of 18 and 27 years old people will be divided into the same category. Therefore, according to the classification results, it is impossible to provide an accurate basis for subsequent research work. Researchers also improve previous methods in a fine-grained way to classify ages [40]. Then, we use this post-sequence dataset to train the methods of supervised learning. Users are divided into different age groups, and other features such as life stages are used for the prediction of fine-grained age.

In summary, most studies on age prediction focus on using the social network data to predict the network users’ ages, while very little research has been conducted on predicting the scholars’ ages. Scholars are a relatively special community for their tremendous contributions to our society, and their ages play significant roles in many important issues such as discovering rising stars and award or funding applications. Therefore, it is very necessary to predict scholars’ ages. In addition, the dataset used in current research is assumed to be complete. However, considering one important particularity of scholars is that they tend to publish articles continuously, which leads to a practical issue that no dataset can be completely accurate. Therefore, inspired by the abovementioned shortcomings of current research, we propose a novel scholars’ age prediction method and the specific procedure is introduced in the following sections.

3. Prediction of Scholar’s Age

Unlike the traditional social network-based scholar attribute mining issues, the types of data obtained from scholars’
publications are relatively simple. Compared with personal information, academic publications can largely reflect the progress of scholars’ research rather than the information that can reveal scholars themselves. Therefore, it poses great challenges to predict scholars’ ages. In order to solve this problem, we first explore the relationship between scholars through an in-depth investigation of the paper information. By analyzing related publications’ datasets and extracting the various relationships hidden in them, we explore the key factors related to the predication of scholars’ ages. Mining parameters that are closely related to the age of scholars depend largely on the type of data. Therefore, the following sections will describe the specific information contained in the dataset. This section will introduce the proposed method of scholars’ age prediction from the perspective of factors which is significant to the ages of scholars and the corresponding predication algorithms.

3.1. Factors Determining the Prediction of Scholar’s Age.
In this part, we propose a series of factors related to the age of scholars. Considering the different professional backgrounds and application scenarios of scholars, we divide the factors into two categories: intuitive factors and complex factors. The intuitive factors refer to those factors that can be directly obtained or calculated through simple operations from the publication’s information while the factors that should be obtained by complex calculations with specialized backgrounds are defined as complex factors in our work. The specific calculation process of these factors is shown as follows.

3.1.1. Factors Can Be Directly Obtained from the Publication’s Information. We first start with the most intuitive information that can be acquired from a publication and utilize it to propose attributes related to the true age of scholars. In this paper, we define the intuitive attributes or factors as follows: the attributes that can be obtained directly from scholars’ publications without complicated operations. The following is a detailed introduction to these intuitive factors that can be obtained by the corresponding papers’ information of scholars.

When giving a list of scholars’ papers, we can intuitively obtain the following information: the number of published papers $N_{pub}$, the total number of papers cited $N_{cit}$, the number of papers published by the first author $N_{AFP}$, the time of scholars to publish the first paper $T_{first}$, the time of scholars to publish the last paper $T_{last}$, the time of publishing the first author paper $T_{Basic}$, the number of collaborators $N_{Coauthor}$, the paper type $AT$, the journal impact factor $JIF$, the total paper length $N_{pp}$, the number of references $N_{ref}$, the funding type $TYPE$, and the funding grant amount $N_{F}$. The abovementioned attributes can be obtained by simply averaging and summing operations.

With the total number of papers published by scholars and the number of papers cited by the scholars, the average reference amount $Ave_{cit}$ of scholars can be obtained. Scholars’ academic age $AA$ can be calculated by the time when the first paper is published and the time when the last paper is published. With the time of the first author’s paper, the academic age $AA_{AF}$ of the scholar as the first author can be calculated. According to the number of papers, the number of first author papers, and the number of collaborators, the average number of collaborators per scholar and the first author paper can be calculated ($N_{Coauthor}$, $N_{AF}$).

On the basis of the number of published papers, the number of first author papers, and the number of references, the average number of references per scholar and first author paper can be obtained ($N_{Ref}$, $N_{AF}^{ref}$). According to the number of scholars’ first author papers and the corresponding citations, the average citation of the first author’s thesis $N_{AF}^{cit}$ can be calculated. By obtaining the scholar’s academic age and the number of papers published, the number of first author papers can be calculated. We set the average number of papers published by scholars each year and the average number of first author paper as $Ave_{Npub}$ and $Ave_{AF}^{NF}$.

Then, according to the total number of papers, the number of first authors, and the length of the papers, the average length of the scholar’s paper $Ave_{Npp}$ and the average length of the first author’s paper $Ave_{AF}^{NF}$ can be obtained. Based on the total number of papers, the first author’s number of papers, and the type of paper, the proportion of research article $Rate_{R}$ and the review article $Rate_{S}$ can be calculated. With the total number of papers, the number of first author papers, and the type of fund funding, the average funded type of each first author’s paper ($Ave_{TYPE}$, $Ave_{AF}^{TYPE}$) can be acquired. Through the total number of papers, the number of papers by the first author, and the number of funds funded, the average amount of funding for each paper and each first author’s paper ($Ave_{NF}$, $Ave_{AF}^{NF}$) can be calculated. The abovementioned intuitive factors can be summarized by Table 1.

3.1.2. Factors Obtained by Complex Computations. According to the intuitive properties obtained in the previous section, we further introduce the complex properties. Considering the different application scenarios and the expertise of different researchers, we divide the factors that influence the ages of scholars into two categories. Compared with the intuitive attributes which can be obtained through simple mathematical operations, the complex attributes have higher requirements for scholars’ expertise and technologies. Furthermore, for the intuitive attributes, due to its lower dependency of expertise, the application of it is more general than the complex ones. However, the complex attributes can reveal the factors affecting the ages of scholars from more sides even though it results in a higher computation complexity. We will make a detailed description of the process of calculating each complex attribute.

Generally speaking, the professional knowledge of scholars will accumulate with time, as well as their influence. Therefore, we first propose factors describing the academic level of scholars to explore the influence on age predication. There are many ways to evaluate a scholar’s ability, such as the most intuitive citations. Secondly, the scholar’s $h$-index value and its PageRank value in the cooperative network can reflect the scholar’s own ability as well. Meanwhile, the trend
of the scholar’s own $h$-index value can also describe the scholar’s ability from the perspective of the potential of scholars. Inspired by the concept of acceleration in physics, we propose a method to compute scholars’ academic ability acceleration from the perspective of the potential of the scholar’s own $h$-index value. Thus, based on the three above parameters, we can get the intuitive parameters used to measure collaborators’ abilities. These parameters include total number of papers, total citations, and average citations of collaborators as the most intuitive parameters used to measure collaborators’ abilities.

In addition, the importance score of papers in the network can also be taken as a measurement of its quality. We first perform the PageRank algorithm in the citation network to calculate the importance of scholarly papers and the first author papers (PR$_p$, PR$_{AF}$). To some extent, author’s ability can largely reflect the quality of the paper. Subsequently, we measure the impact of the collaborator’s ability and their backgrounds to reveal the quality of the paper. Three parameters including total number of papers, total citations, and average citations of collaborators are the most intuitive parameters used to measure collaborators’ abilities. Then, based on the three above parameters, we can get the sum of the partner’s $h$-index value, the largest $h$-index value of the collaborators, the smallest $h$-index value of the collaborators, and the partner’s average $h$-index value. Furthermore, we will obtain the difference between scholar’s and the corresponding $h$-index value.

Except for evaluating the influence of collaborators, the differences between the research background and the scholars themselves are used to measure the quality of the paper. We describe the differences between scholars and their collaborators in two ways. One consideration is institutional information and national differences, and the

### Table 1: Intuitive factor descriptions.

| Feature | Description |
|---------|-------------|
| $N_{pub}$ | The number of publications |
| $N_{at}$ | The total citation counts |
| $N_{AFP}$ | The number of publications as the first author |
| $T_{first}$ | The publication time of the first paper |
| $T_{last}$ | The publication time of the last paper |
| $T_{AdOpt}$ | The publication time of the first paper as the first author |
| $N_{coauthor}$ | The number of co-authors |
| $AT$ | The number of foundations |
| $JIF$ | The average of citation counts |
| $N_{PP}$ | The academic age of scholar |
| $N_{ref}$ | The academic age of scholar as the first author |
| $TYPE_p$ | The journal impact factor |
| $N_F$ | The proportion of research article type |
| Ave$_{cit}$ | The average number of foundations |
| AA | The average of citation counts of first author papers |
| $AA_{AF}$ | The average reference numbers per first author paper |
| $N_{coauthor}$ | The average reference numbers per paper |
| $N_{ref}$ | The number of references |
| $N_{AF}$ | The number of co-authors per paper |
| $N_{AFP}$ | The number of publications as the first author |
| Ave$_{Npub}$ | The average number of the first author papers per year |
| Ave$_{NPP}$ | The average number of papers per year |
| Ave$_{AF}$ | The average number of papers per year |
| $N_{AF}$ | The average number of the first author papers |
| Rate$_F$ | The average number of foundations |
| Rate$_S$ | The average number of foundations types |
| Ave$_{NF}$ | The average number of foundations of first author papers |
| Ave$_{TYPEF}$ | The average number of foundation types of first author papers |
| Ave$_{AF}$ | The average number of foundations of first author papers |

In the above formula, we define scholars’ academic ability acceleration as follows:

$$A = \frac{V(t + \Delta t) - V(t)}{\Delta t},$$  \hspace{1cm} (1)

where $A$ represents the acceleration in $\Delta t$ time, $V(t + \Delta t)$ is the speed at time $t + \Delta t$, and $V(t)$ is the speed at time $t$.

On the basis of the above formula, we define scholars’ academic ability acceleration as follows:

$$V(a_i) = \frac{h(a_i)^{(t + \Delta t)} - h(a_i)^{(t)}}{\Delta t},$$

$$A(a_i) = \frac{V(a_i)^{(t)} - V(a_i)^{(t - \Delta t)}}{\Delta t},$$ \hspace{1cm} (2)

where $V(a_i)$ represents the speed at which scholars’ influence varies in time $\Delta t$, $h(a_i)^{(t)}$ is the scholar’s $h$-index value at time $t$ and $h(a_i)^{(t + \Delta t)}$ is the $h$-index value of the scholar at time $t + \Delta t$. $A(a_i)$ indicates the acceleration of scholars’ influence in time $\Delta t$, $V(a_i)^{(t - \Delta t)}$ is the speed of scholar’s influence variation at time $t - \Delta t$, and $V(a_i)^{(t)}$ represents the speed at which the scholar’s influence at time $t$.

In addition, the importance score of papers in the network can also be taken as a measurement of its quality. We first perform the PageRank algorithm in the citation network to calculate the importance of scholarly papers and the first author papers (PR$_p$, PR$_{AF}$). To some extent, author’s ability can largely reflect the quality of the paper. Subsequently, we measure the impact of the collaborator’s ability and their backgrounds to reveal the quality of the paper. Three parameters including total number of papers, total citations, and average citations of collaborators are the most intuitive parameters used to measure collaborators’ abilities.
other is the difference in the direction of scholars’ research. We take the information of keywords in papers as an index to represent the research direction of scholars themselves. Specifically, the difference between collaborators can be measured by their information entropy. Furthermore, the difference between scholars’ research directions is taken into account. The specific calculation process is as follows:

\[
\text{Div}(a_i)_{\text{inst}} = -\sum_{m=1}^{r} w_m \log_2 (w_m), \\
\text{Div}(a_i)_{\text{out}} = -\sum_{n=1}^{q} y_n \log_2 (y_n), \\
\text{Div}(a_i)_{\text{key}} = -\sum_{p=1}^{q} k_p \log_2 (k_p),
\]

where \( \text{Div}(a_i)_{\text{inst}}, \text{Div}(a_i)_{\text{out}}, \) and \( \text{Div}(a_i)_{\text{key}} \) represent scholar \( a_i \)'s difference between their collaborators in the organization, country, and research direction. \( w_m \) is the frequency at which the word \( m \) appears in all collaborators’ information, and \( r \) is the total number of words \( m \). \( y_n \) is the frequency of the word \( n \) appearing in the country information of all collaborators, and \( n \) is the total number of words \( n \). \( k_p \) is the frequency of word \( \rho \) in all the keywords information in the collaborator’s paper, and the total number of \( q \) is \( \rho \).

Finally, we assume that the writing style of scholars varies with the growth of their academic age. As scholars enter the academia, their writing skills and the quality of their thesis will be greatly improved, and their writing style will also change. In this work, due to the limitations of datasets, instead of the original text, we analyze the abstract information to extract the features related to a scholar’s writing skills. Considering the efficiency and accuracy of our framework, we transform the textual abstract data into a low-dimensional space. The representation learning algorithm can automatically learn the representation of such abstract information. In natural language processing, representation learning can extract the intrinsic features hidden in the words, paragraphs, or chapters. One superiority of representation learning is that it can transform the data into a low-dimension vector while preserve essential features of raw data. Therefore, according to the algorithm of word representation learning, we represent the abstract information and convert it into a low-dimensional dense vector as the input of the prediction algorithm.

The representation learning algorithm we choose is a word-based representation algorithm called paragraph2vec, which can obtain ideal results in the process of short text data [41, 42], and it can learn the short text information more effectively. The details of this algorithm will be described below. The first part of the algorithm is to represent the words. It maps every word into a matrix \( W \) and the index of element represents the word’s position in the dictionary. Given a list of training phrases \( w_1, w_2, \ldots, w_n \), the goal of the word vector is to maximize the average log probability, which is calculated by the following formula:

\[
\frac{1}{N} \sum_{i=j}^{N} \log p(w_i|w_{i-j}, \ldots, w_{i+j}).
\]

To solve the above formula, we convert it to the following equation based on the softmax method:

\[
p(w_i|w_{i-j}, \ldots, w_{i+j}) = \frac{e^{y_i}}{\sum_{v \in \mathcal{V}} e^{y_v}},
\]

where \( y_v \) is the unnormalized logarithmic probability for each output word \( v \) and can be calculated by the following formula:

\[
y = b + U h(w_{i-j}, \ldots, w_{i+j}; W),
\]

where \( U \) and \( b \) are parameters and \( h \) represents the sum of the column vectors extracted from the matrix \( W \) and matrix \( D \).

3.2. Prediction Method. The detailed calculation procedure of each factor that related to the age of scholars is described in the above. Next, we will elaborate the specific prediction algorithm. The predictive task of this paper can be defined as predicting the true age of scholars through the data extracted from scholarly papers. Suppose \( A = A_1, A_2, A_3, \ldots, A_n \) represents the true age sequence of \( n \) scholars, and \( S = S_1, S_2, S_3, \ldots, S_n \) indicates the features extracted from the dataset representing these scholars. In this paper, we input these features into the corresponding prediction algorithm and obtain an output function \( F = F_1, F_2, F_3, \ldots, F_n \), where \( F_i \) represents the age prediction of scholar \( i \). We will describe the calculation process of the prediction algorithm in detail as shown in Figure 1.

The learning algorithm chosen for this paper is XGBoost. XGBoost is an end-to-end scalable method based on tree promotion, which is an improvement of the GBDT algorithm. Due to its superior efficiency and high accuracy, it has attracted more and more scientists’ attention. The GBDT algorithm consists of multiple decision trees, and the results of all the subtrees are incorporated into together to get the final output. On the basis of the GBDT algorithm, we then describe the XGBoost algorithm in detail below. The first part of the XGBoost algorithm is the regularization of learning objectives. Given a dataset \( D = (x_i, y_i) \) containing \( n \) samples and \( m \) features, the tree integration model using the \( K \) additive functions can predict the output as follows:

\[
\bar{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F,
\]
where $f = f(x) = w^T(x)q \in \mathbb{R}^m \rightarrow T$, $w \in \mathbb{R}^T$ is the vector space of the regression tree, $q$ represents the structure of the tree that maps the sample to the corresponding leaf node, and $T$ is the number of leaf nodes that the tree contains. $f_k$ corresponds to each tree with a structure of $q$ and a leaf weight of $w$. Unlike the decision tree, each regression tree includes consecutive values for each leaf node, and $w_i$ represents the value of the $i$th leaf node. In summary, it is classified into leaves by the decision rule in the tree ($q$), and the values of all corresponding leaves ($w$) are summed to predict the final result. In order to learn the rules in the prediction model, the following regular objective functions are minimized:

$$L(\phi) = \sum_i l\tilde{y}_i, y_i + \sum_k \Omega(f_k),$$

$$\Omega(f) = yT + \frac{1}{2}\lambda\|w\|^2,$$

where $l$ is a cost function and can be used to calculate the prediction result (the difference between $\tilde{y}_i$ and the actual value ($y_i$)) and $\Omega$ is a regular term to avoid overfitting.

The traditional European vector space used in the optimization method cannot be applicable in the tree integration model described above; therefore, the accumulation method is incorporated into the training part. The purpose of adding function $f_t$ is to improve the model:

$$L^{(t)} = \sum_{i=1}^n \left(l\left(y_i, \tilde{y}_i^{-1} + f_i(x_i)\right) + \Omega(f_t)\right),$$

where $\tilde{y}_i$ represents the predicted result of the $i$th instance at $t$ iteration.

Subsequently, the model is optimized quickly by using the second Taylor expansion method as follows:

$$L^{(t)} = \sum_{i=1}^n \left[l\left(y_i, \tilde{y}_i^{-1}\right) + g_t f_i(x_i) + \frac{1}{2}h_t f_t^2(x_i)\right] + \Omega(f_t),$$

where $\Omega$ is a regular term to avoid overfitting.
where \( g_i = \delta_{y_i} l'(y_i, y_i^{(t-1)}) \) and \( h_i = \delta_{y_i}^2 l'(y_i, y_i^{(t-1)}) \) represent the first order and the second order gradient statistics of the cost functions, respectively. In order to simplify the function, the constants can be removed. The simplified formula is as follows:

\[
L^{(t)} = \sum_{i=1}^{n} \left[ g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i) \right] + \Omega(f_i). \tag{11}
\]

Define the sample set of the leaf node \( j \) as the function \( L_j = ilq(x_i) = j \), and then, convert the above formula by expanding the regular item \( \Omega \) as follows:

\[
\bar{L}^{(t)} = \sum_{i=1}^{n} \left[ g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i) \right] + y^T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2
\]

\[
= \sum_{j=1}^{T} \left( \sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \sum_{i \in I_j} h_i + \lambda \right) w_j^2 + y^T. \tag{12}
\]

For a given tree structure \( q(x) \), the optimization process for the weight of the leaf node \( j \) \( w_j^* \) is as follows:

\[
w_j^* = \frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}. \tag{13}
\]

Based on this, the corresponding optimization function can be obtained. The calculation process is as follows:

\[
\bar{L}^{(t)}(q) = \frac{1}{2} \sum_{j=1}^{T} \left( \sum_{i \in I_j} g_i \right)^2 + y^T. \tag{14}
\]

The above formula can be used to evaluate the score of a tree structure \( q \).

In summary, based on the factors we proposed, we predict the ages of scholars by the XGBoost algorithm. More importantly, we analyze the importance and influence of each feature and then calculate the score of each feature. The validity of our method in predicting the age of scholars will be verified below.

### 4. Experiments and Results

In this section, we use the Web of Science dataset to evaluate the performance of the proposed method. Firstly, we investigate the accuracy of predicting the age of scholars between each machine learning algorithm and comparison method by the factors we proposed. Then, the importance scores of the abovementioned factors are discussed. Finally, the prediction accuracy of each method with the incomplete datasets is verified. We will first introduce the datasets used in this article and the comparison methods in the following parts.

**4.1. Dataset.** In order to predict the true ages of scholars, it is necessary to analyze the papers of scholars. Since the age belongs to the privacy information, it is difficult to obtain.

Furthermore, the same name of two scholars is an inevitable problem in experiment. Based on this, we choose the Nobel Prize winner and the Turing Award winner data as our dataset. These two rewards publish the winners’ specific year of birth in the official website which can eliminate errors as far as possible.

Since we focus on the scholarly papers, we remove the Nobel Peace Prize and the Literature Award due to the specificity of them. The scholars’ data in this section come from the Turing Award winner, the Nobel Prize in Chemistry, the Nobel Prize in Physics, the Nobel Prize in Physiology or Medicine, and the Nobel Prize in Economics. Due to the diversity of academic disciplines, we obtained datasets for experiments from the Web of Science. We will introduce the details of the Nobel Prize winners and the Turing Award winners, as shown in Table 3.

We then describe the information contained in the Web of Science dataset. Web of Science is an online digital library that contains a large amount of paper data from various disciplines. The paper information of Web of Science is updated constantly. This information mainly includes topics, authors, abstracts, institutions, citations, keywords, journals, dates, paper types, and funding information. It can be seen that the information provided by Web of Science is comprehensive. Based on the Nobel Prize and the Turing Award winners’ lists, we crawl their corresponding papers from the Web of Science database and the final subdataset contains 486 scholars and 38,478 papers.

**4.2. Baseline Methods.** To investigate the effectiveness of our work, we compare it with the following baseline methods:

(a) GBDT algorithm: it is suitable for dealing with the nonlinear relationship between a large number of predicted feature variables and target variables. It is an iterative decision tree algorithm. The algorithm consists of multiple decision trees, and the results of all trees are added up as the final output.

(b) SVR: SVR is a support vector machine (SVM)-based regression model. The basic framework of SVR is similar to SVM. The difference between them is that SVM is aiming at solving the classification problems while SVR is aiming at solving the regression problems.

**4.3. Evaluation Metrics.** To evaluate the performance of different learning algorithms and factors, four typical metrics are put forward: MAE (mean absolute error), MAPE (mean absolute error percentage), MSE (mean squared error), ACC (accuracy), and \( R^2 \). Given the true value \( y \) and the predicted value \( \tilde{y} \), the values of the above various evaluation indicators can be calculated.

**4.4. Prediction of Scholar’s Age.** In this paper, we classify the influencing factors into intuitive factors and complex factors. The main purpose of this paper is to consider the research background and application scenarios of different...
of scholars’ ages. In summary, the XGBoost algorithm achieves the highest accuracy in all comparison methods, and the prediction result of the XGBoost\textsuperscript{SC} method is the best.

Subsequently, we use the above evaluation metrics MSE, MAE, MAPE, ACC, and \( R^2 \) to further measure the effectiveness of the prediction results of each method. Among them, MSE, MAE, and MAPE are used to measure the difference between the predicted result and the true value. The smaller their values are, the better the predicted performance is. Conversely, \( R^2 \) represents the correlation between the predicted result and the true value, and ACC can show the accuracy of the predicted result, so the larger the values of both are, the better the prediction accuracy is. According to Table 4, using XGBoost as the prediction algorithm still achieves the best performance. Among them, under the same condition, XGBoost\textsuperscript{3}, XGBoost\textsuperscript{SC}, and XGBoost\textsuperscript{SC} are more accurate than other prediction algorithms. At the same time, the accuracy of using all factors to predict the age of scholars is still the highest among all methods. The accuracy of the XGBoost\textsuperscript{SC} algorithm exceeds 90%, and its \( R^2 \) value also represents a high degree of correlation for predicting scholars’ ages.

### 4.5. Factor Contribution Analysis

Through the above analysis, the accuracy of each method for scholars’ age prediction can be verified, while the importance and contribution of each factor remains to be explored. Firstly, the importance of each feature is calculated by using the feature importance in the machine learning algorithm. Then, the jackknife method is used to evaluate the contribution of both intuitive factors and complex factors. The jackknife method consists of two cases: (1) using only one factor to predict (add); (2) removing a set of factors and using the remaining factors for prediction (subtraction). Based on these two situations, the individual contributions of the above two types of factors to the overall prediction task can be explored.

According to the above experimental results, the accuracy of the scholars’ age prediction is the highest when the intuitive attributes and complex attributes are combined by utilizing the XGBoost algorithm. Therefore, we mainly analyze the importance of each factor in the XGBoost\textsuperscript{SC} method. In the XGBoost algorithm, the importance of a feature is equal to the number of nodes whose decision tree splits. Figures 3 and 4 show the importance of all factors. As seen from the abovementioned results, the overall importance scores of the intuitive factors are higher than the complex factors, and the time when the scholar publishes the first paper is the most important feature. Specifically, the top

### Table 3: Statistical information of the dataset.

| Discipline                  | Time span | Number of scholars | Number of papers |
|-----------------------------|-----------|--------------------|------------------|
| Computer science            | 1915–1952 | 65                 | 5804             |
| Physics                     | 1845–1974 | 136                | 7993             |
| Chemistry                   | 1852–1960 | 134                | 11660            |
| Physiology or medicine      | 1843–1963 | 176                | 10629            |
| Economics                   | 1895–1953 | 72                 | 2392             |

For example, when researchers cannot get more information to calculate the complex attributes, they can make simple predictions about the ages of scholars by using intuitive factors. On the contrary, when researchers aim to obtain higher prediction accuracy, they can consider intuitive and complex factors simultaneously. Because the probability of scholars who have the same birth year is relatively small, we set a space strip with a width of 2\( \varepsilon \). When the predicted result falls into this interval, the prediction result is considered to be accurate.

The accuracy of each method in predicting scholars’ age is explored at first. Among them, GBDT\textsuperscript{3}, XGBoost\textsuperscript{3}, and SVR\textsuperscript{3} represent the performance of using only the intuitive factors under each prediction algorithm; GBDT\textsuperscript{SC}, XGBoost\textsuperscript{SC}, and SVR\textsuperscript{SC} show the performance of using complex factors under various prediction algorithms; GBDT\textsuperscript{SC}, XGBoost\textsuperscript{SC}, and SVR\textsuperscript{SC} indicate the predictive performance of utilizing both the intuitive and complex factors at the same time to predict the ages of scholars. The interval with a width of 2\( \varepsilon \) consists of \( y = x + 2 \) and \( y = x - 2 \) can be seen in these figures. The closer the prediction results to \( y = x \) are, the better the prediction accuracy will be. It can be seen from Figure 2 that when the intuitive and the complex factors are used simultaneously to predict the ages of scholars, the performance is the best. Specifically, when using intuitive attributes, the performance of the GBDT\textsuperscript{SC}, XGBoost\textsuperscript{SC}, and SVR\textsuperscript{SC} can obtain the highest prediction accuracy. Although all attributes can achieve the highest accuracy, their correlation for predicting scholars’ ages. When researchers do not pursue higher accuracy, using intuitive factors for prediction alone can simplify the calculation process and improve the efficiency of the prediction.

When the scholar’s birth year is between 1920 and 1960, it can be observed that the prediction results in this interval are better. If a scholar’s birth year is not in this interval, there are more outliers. This is mainly because of the incompleteness of datasets. Scholars born before 1920 and after 1960 are accounted for 26% of the total number of scholars. However, the number of papers published by these scholars only accounts for 14%. This is because early scholars are limited by the way of acquiring knowledge and the difficulty of communication between scholars. Therefore, the frequency of cooperation among scholars is less than the scholars’ today. Thus, the data of early scholars in the database are incomplete and result in errors in the prediction.
ten features are $T_{first}$, $T_{last}$, $T_{first}$, $N_{coauthor}$, $AA$, $AA^{AF}$, $N^{AF}_{coauthor}$, $h$–index, $Q(a_i)$, and $A_a$. According to their rankings, the academic age of scholars is also very important for predicting the true ages of scholars, and the number of papers and the influence of collaborators are also crucial for predicting their ages. Different from previous studies, we

![Graphs showing prediction results by SVR, GBDT, and XGBoost.](image)

**Figure 2**: Prediction results by SVR, GBDT, and XGBoost. (a) SVR$^S$. (b) SVR$^C$. (c) SVR$^{SC}$. (d) GBDT$^S$. (e) GBDT$^C$. (f) GBDT$^{SC}$. (g) XGBoost. (h) XGBoost$^C$. (i) XGBoost$^{SC}$.

**Table 4**: Prediction results of each method.

| Method      | MAE   | MAPE  | MSE   | ACC   | $R^2$  |
|-------------|-------|-------|-------|-------|--------|
| SVR$^S$     | 1.981 | 0.476 | 3.712 | 0.291 | 0.663  |
| SVR$^C$     | 2.275 | 0.693 | 3.819 | 0.251 | 0.527  |
| SVR$^{SC}$  | 1.336 | 0.312 | 2.836 | 0.514 | 0.737  |
| GBDT$^S$    | 1.056 | 0.275 | 2.487 | 0.696 | 0.792  |
| GBDT$^C$    | 0.915 | 0.289 | 2.628 | 0.713 | 0.726  |
| GBDT$^{SC}$ | 0.816 | 0.194 | 2.364 | 0.791 | 0.832  |
| XGBoost$^S$ | 0.732 | 0.190 | 1.228 | 0.803 | 0.964  |
| XGBoost$^C$ | 0.746 | 0.215 | 1.363 | 0.824 | 0.902  |
| XGBoost$^{SC}$ | 0.689 | 0.071 | 1.046 | 0.903 | 0.972  |
also consider the impact of research funds in this paper. The experimental results show that the types of research funds also have a certain impact on the prediction of scholars’ ages. Subsequently, the jackknife method is applied to perform the contribution analysis of these two types of factors. As shown in Figure 5, when the intuitive factor is removed, the accuracy of the obvious prediction result is significantly reduced. This result demonstrates the importance of intuitive attributes in predicting the ages of scholars. However, when the complex factor is removed, the accuracy is degraded slightly. This result also proves the importance of intuitive factors on the other side. When the factors’ types are increased, the obvious improvement of the accuracy rate after the addition of the intuitive factors can also be observed. This paper verifies that the intuitive factor plays an important role in predicting the scholar’s age from multiple angles. In summary, when researchers do not have higher accuracy requirements for scholars’ age, they can use the intuitive factors to have a quick estimation of the scholar’s age range.

4.6. Prediction Results under Incomplete Dataset. Due to the dynamic nature of academic data, there actually exists no dataset that accurately contains all the paper information. Therefore, the incomplete data will affect the results of the experiment. The complete dataset in this work refers to the whole publications’ data of a scholar we acquired from the WOS database. And the incomplete dataset refers to the missing of several articles of scholars. In order to explore the impact of incomplete dataset on the experimental results, we randomly remove the 10% and 30% of the papers of each scholar in the dataset and then predict the ages of scholars.

It can be seen from Figure 6 that when the dataset is removed by 10%, the prediction accuracy of each method is reduced, but the difference is not obvious. To further
Figure 5: Factor contribution analysis. Four models trained with only or without the denoted factors. F, full feature set; I, intuitive factors; C, complex factors. (a) SVR. (b) GBDT. (c) XGBoost.

Figure 6: The prediction results under 10% loss of data by SVR, GBDT, and XGBoost. (a) XGBoost. (b) XGBoost. (c) XGBoost. (d) SVR. (e) SVR. (f) SVR. (g) GBDT. (h) GBDT. (i) GBDT.
demonstrate the experimental results, we calculated the MSE, MAE, MAPE, ACC, and $R^2$ values for each method. As shown in Table 5, the accuracy of each method is decreased slightly with the incomplete dataset. Therefore, when there is a small amount of missing data in the datasets, the impact on the experimental results is not significant. Subsequently, the

| Method | MAE   | MAPE  | MSE    | ACC   | $R^2$ |
|--------|-------|-------|--------|-------|-------|
| SVR$^S$ | 2.382 | 0.635 | 3.937  | 0.274 | 0.583 |
| SVR$^C$ | 2.572 | 0.728 | 4.193  | 0.228 | 0.517 |
| SVR$^{SC}$ | 1.963 | 0.516 | 3.173  | 0.461 | 0.613 |
| GBDT$^S$ | 1.825 | 0.614 | 3.452  | 0.423 | 0.592 |
| GBDT$^C$ | 1.615 | 0.624 | 2.941  | 0.691 | 0.548 |
| GBDT$^{SC}$ | 1.116 | 0.536 | 2.803  | 0.706 | 0.738 |
| XGBoost$^S$ | 0.794 | 0.284 | 1.938  | 0.791 | 0.735 |
| XGBoost$^C$ | 0.966 | 0.315 | 2.376  | 0.738 | 0.638 |
| XGBoost$^{SC}$ | 0.744 | 0.241 | 1.847  | 0.849 | 0.794 |

Figure 7: The prediction results under 30% loss of data by SVR, GBDT, and XGBoost. (a) XGBoost$^S$. (b) XGBoost$^C$. (c) XGBoost$^{SC}$. (d) SVR$^S$. (e) SVR$^C$. (f) SVR$^{SC}$. (g) GBDT$^S$. (h) GBDT$^C$. (i) GBDT$^{SC}$. 

Table 5: Prediction results under 10% loss of data.
data are then removed by 30%. As shown in Figure 7 and Table 6, the accuracy is significantly reduced in the case of missing 30% data compared to the prediction result using the complete dataset. It can be inferred that the large-scale missing data will cause a serious impact on both the experimental results and the verification of the method. Among all the methods, the prediction accuracy of the XGBoost algorithm is still the highest.

5. Conclusion

Related studies have shown that scholar’s actual age plays a crucial role in evaluating and predicting their scientific impact, and because of the privacy issues, the age data are difficult to obtain directly. Inspired by this fact, our main purpose in this paper is to predict the ages of scholars through scholarly articles’ information. In order to solve this problem, we first explore the factors that affect the ages of scholars. Considering the different research backgrounds of the researchers and different application scenarios, the influencing factors are divided into two categories according to the computational complexity, which are the intuitive factors and complex factors. Then, we transform the textual data into a low-dimensional vector by the representation learning method, and the machine learning algorithm is utilized to predict the scholar’s age.

Due to the difficulty of acquiring the true age of scholars and the name ambiguity problem, we crawl the age data of Nobel and Turing Award winners to verify the validity of our method. Based on the list of winners in Nobel Prize in Chemistry, the Nobel Prize in Physiology or Medicine, the Nobel Prize in Economics, and the Turing Award winners, their papers are correspondingly obtained from the Web of Science database. The experimental results show that the accuracy of the prediction of scholars’ age is above 90% by using the method proposed in this paper, and the combination of intuitive factors and complex factors shows the best performance when predicting scholar’s age. Meanwhile, the accuracy of using intuitive factors alone is higher than that of complex factors. Among all the factors, the time when scholars published the first paper, the time of the last paper, and the influence of their collaborators are the most relative factors. Furthermore, the performance of our method is still the best among all the baselines in spite of the incomplete dataset. In the future, we will conduct experiments on a wider variety of datasets from more disciplines to prove the effectiveness of our work.

Data Availability

The Nobel Prize and the Turing Award winners’ specific year of birth can be acquired from their official websites (https://www.nobelprize.org/ and https://amturing.acm.org/byyear.cfm). Based on the Nobel Prize and the Turing Award winners’ lists, we access their corresponding papers’ information from the Web of Science database (http://www.webofknowledge.com/). The publication data used to support the findings of this study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

The authors would like to thank everyone in the Alpha Lab of Dalian University of Technology especially Weixin Shi for the valuable contributions in the process of writing this article. This work was supported by the Natural Science Foundation of China under grant nos. 71904022 and 62072067 and the Fundamental Research Funds for the Central Universities under grant no. DUT20RC(4)026.

References

[1] J. Liu, J. Tian, X. Kong, I. Lee, and F. Xia, “Two decades of information systems: a bibliometric review,” Scientometrics, vol. 23, 2018.
[2] J. Liu, X. Kong, X. Feng et al., “Artificial intelligence in the 21st century,” IEEE Access, vol. 6, no. 99, 2018.
[3] A. Clauset, D. B. Larremore, and R. Sinatra, “Data-driven predictions in the science of science,” Science, vol. 355, no. 6324, pp. 477–480, 2017.
[4] F. Xia, W. Wang, T. M. Bekele, and H. Liu, “Big scholarly data: a survey,” IEEE Transactions on Big Data, vol. 3, no. 1, pp. 18–35, 2017.
[5] J. Zhang, Y. Hu, Z. Ning et al., “Author impact ranking through positions in collaboration networks,” 2018.
[6] B. F. Jones, “The burden of knowledge and the “death of the renaissance man”: is innovation getting harder?” Review of Economic Studies, vol. 76, no. 1, pp. 283–317, 2009.
[7] E. Balsa, C. Perez-Solà, and C. Díaz, “Towards inferring communication patterns in online social networks,” ACM Transactions on Internet Technology, vol. 17, no. 3, pp. 1–21, 2017.
[8] P. Shukla and J. Drennan, “Interactive effects of individual- and group-level variables on virtual purchase behavior in online communities,” Information & Management, vol. 55, no. 5, pp. 598–607, 2018.
[9] A. Alekseev and S. I. Nikolonenko, “Predicting the age of social network users from user-generated texts with word embeddings,” 2017.
[10] P. Wang, J. Guo, Y. Lan, J. Xu, and X. Cheng, “Your cart tells you: inferring demographic attributes from purchase data,” 2017.
[11] J. Qian, X. Y. Li, C. Zhang, and L. Chen, "De-anonymizing social networks and inferring private attributes using knowledge graphs," 2016.

[12] X. Lin, M. Featherman, and S. Sarker, "Understanding factors affecting users' social networking site continuance: a gender difference perspective," *Information & Management*, vol. 54, no. 3, pp. 383–395, 2017.

[13] P. Wang, F. Sun, D. Wang, J. Tao, X. Guan, and A. Bifet, "Inferring demographics and social networks of mobile device users on campus from ap-trajectories," 2017.

[14] R. Y. Dougnon, P. Fournier-Viger, J. C.-W. Lin, and R. Nkambou, "Inferring social network user profiles using a partial social graph," *Journal of Intelligent Information Systems*, vol. 47, no. 2, pp. 313–344, 2016.

[15] J. Jia, B. Wang, L. Zhang, and N. Z. Gong, "Attrinfier: Inferring user attributes in online social networks using markov random fields," 2017.

[16] W. Wang, S. Yu, T. M. Bekele, X. Kong, and F. Xia, "Scientific collaboration patterns vary with scholars' academic ages," *Sociometrics*, vol. 11, no. 1, pp. 329–343, 2017.

[17] S. Mukherjee, D. M. Romero, B. Jones, and B. Uzzi, ""The nearly universal link between the age of past knowledge and tomorrow' breakthroughs in science and technology: the hotspot," *Science Advances*, vol. 3, no. 4, 2017.

[18] R. Y. Dougnon, P. Fournier-Viger, J. C.-W. Lin, and R. Nkambou, *Accurate Online Social Network User Profiling*, Springer International Publishing, Berlin, Germany, 2015.

[19] J. Chen, Y. Liu, and M. Zou, "Home location profiling for users in social media," *Information & Management*, vol. 53, no. 4, pp. 135–143, 2016.

[20] N. Garera and D. Yarowsky, "Structural, transitive and latent models for biographic fact extraction," 2009.

[21] G. S. Mann and D. Yarowsky, "Multi-field information extraction and cross-document fusion," 2005.

[22] S. Bergsma, M. Dredze, B. Van Durme, T. Wilson, and D. Yarowsky, "Broadly improving user classification via communication-based name and location clustering on twitter," 2013.

[23] Q. Fang, J. Sang, C. Xu, and M. S. Hossain, "Relational user attribute inference in social media," *IEEE Transactions on Multimedia*, vol. 17, no. 7, pp. 1031–1044, 2015.

[24] E. Zheleva and L. Getoor, "To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles," 2009.

[25] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel, "You are who you know:inferring user profiles in online social networks," 2010.

[26] X. Cao, Y. Chen, C. Jiang, and K. J. R. Liu, "Evolutionary information diffusion over heterogeneous social networks," *IEEE Transactions on Signal & Information Processing Over Networks*, vol. 2, no. 4, pp. 595–610, 2016.

[27] N. Garera and D. Yarowsky, "Modeling latent biographic attributes in conversational genres," 2009.

[28] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in twitter," 2010.

[29] S. Volkova and B. Van Durme, "Online bayesian models for personal analytics in social media," 2015.

[30] A. Culotta, N. K. Ravi, and J. Cutler, "Predicting the demographics of twitter users from website traffic data," 2015.

[31] D. Rao, M. Paul, C. Fink, D. Yarowsky, T. Oates, and G. Coppersmith, "Hierarchical bayesian models for latent attribute detection in social media," 2011.

[32] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing, "A latent variable model for geographic lexical variation," 2010.

[33] X. Sun, J. Guo, X. Ding, and T. Liu, "A general framework for content-enhanced network representation learning," 2016.

[34] D. Nguyen, D. Trieschnigg, A. S. Doğruöz et al., "Why gender and age prediction from tweets is hard: lessons from a crowdsourcing experiment," *Association for Computational Linguistics*, vol. 2014, 2014.

[35] C. Yang, D. Zhao, D. Zhao, E. Y. Chang, and E. Y. Chang, "Network representation learning with rich text information," 2015.

[36] J. Hu, H. J. Zeng, H. Li, C. Niu, and Z. Chen, "Demographic prediction based on user's browsing behavior," 2007.

[37] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, "A probabilistic fusion approach to human age prediction," 2008.

[38] N. Dong and N. A. Smith, "Author age prediction from text using linear regression," 2011.

[39] G. F. Nane, V. Lariviére, and R. Costas, "Predicting the age of researchers using bibliometric data," *Journal of Informetrics*, vol. 11, no. 3, pp. 713–729, 2017.

[40] D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder, "‘How old do you think i am?’; a study of language and age in twitter," 2013.

[41] X. Rong, "Word2vec parameter learning explained," 2020.

[42] Y. Zhang, D. Shen, G. Wang, Z. Gan, R. Henao, and L. Carin, "Deconvolutional paragraph representation learning," 2017.