World Wide Web (2023) 26:187–205
https://doi.org/10.1007/s11280-022-01010-9

A multi-attribute decision making approach based on information extraction for real estate buyer profiling

Linan Zhu[1] · Minhao Xu[1] · Yifei Xu[1] · Zhechao Zhu[1] · Yanyan Zhao[2] · Xiangjie Kong[1]

Received: 20 October 2021 / Revised: 21 December 2021 / Accepted: 10 January 2022 /
Published online: 9 February 2022

Abstract
With the rapid development of the Internet and the widespread usage of mobile terminals, data-driven user profiling has become possible. User profiles describe the user's overall behavior characteristic from multiple perspectives (e.g. basic information, feature preference, social attribute), which can explore the potential relationships between complex user behaviors and the decision-making process. In this paper, we focus on the problem of real estate buyer profiling and propose a novel multi-attribute decision making (MADM) approach, trying to solve the needs of enterprises to locate target customers accurately. Firstly, we reorganize the dataset by integrating structured with unstructured data, where an **E**nriched **B**i-directional long short-term memory (Bi-LSTM) **C**onditional **R**andom **F**ield (EB-CRF) model is proposed to extract important information in the unstructured data. Based on four general dimensions (i.e. basic information, family situation, purchase intention, financial situation), we then design an entropy-based weight allocation algorithm to obtain attribute weights, which helps explore implicit heterogeneous relationships. Finally, with the help of expert knowledge, we use attribute weights and representation technology "bag of attributes" to construct a buyer-specific feature representation. Extensive experimental results indicate that our approach outperforms strong baselines significantly and achieves state-of-the-art performance.

Keywords Multi-attribute decision making · Real estate buyer profiling · Information extraction · Representation learning · Heterogeneous relationship

This article belongs to the Topical Collection: *Special Issue on Decision Making in Heterogeneous Network Data Scenarios and Applications*
Guest Editors: Jianxin Li, Chengfei Liu, Ziyu Guan, and Yinghui Wu

✉ Xiangjie Kong
xjkong@ieee.org

[1] College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China

[2] College of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

1 Introduction

Decision making is a complex thought operation process, which involves problem identification, information collection, alternatives evaluation and ultimately drawing conclusions [30]. During the decision-making process, uncertainty and complexity are inevitable under human factors, how to deal with massive information accessed from various sources of the Internet is becoming more and more important.

With the growing maturity of big data technology and natural language processing (NLP) technology, data-driven user profiling has become possible to address this issue. Based on the behavioral characteristics of users in real life, user profiles abstract labels from multiple dimensions (e.g. basic information, feature preference, social attribute) and aim to describe the user's overall behavioral characteristics as comprehensively as possible. Therefore, user profiles can explore implicit heterogeneous relationships and help provide quality services, which can be applied in many domains.

In particular, the real estate industry is a domain in urgent need of user profiling. The real estate industry plays a decisive role in China's economic development [17]. In the past, real estate enterprises usually carried out promotion through large-scale advertising, questionnaire surveys and telephone interviews. However, these methods have the disadvantage of large investment and inconspicuous effects. Specifically, enterprises contact customers actively and make customers accept passively, which hinders the long-term development of enterprises. To address the above issues, user profiles describe the overall behavioral characteristics of users by collecting massive user information, which can help enterprises locate target customer groups and conduct personalized precision marketing for customers with demands. In this way, both enterprises and customers can achieve a win-win situation.

Reviewing the previous literature, there are rare real estate buyer profiling researches. Therefore, this paper mainly conducts groundbreaking work. In this paper, we propose a novel multi-attribute decision making approach (MADM) for real estate buyer profiling. Firstly, we reorganize the dataset by integrating structured with unstructured data. Specially, we use an **E**nriched **B**i-directional long short-term memory (Bi-LSTM) **C**onditional **R**andom **F**ield (EB-CRF) model and man-made templates to obtain structured data from the unstructured data (e.g. descriptive text), which can fill in the missing values of the original structured data. Based on four general dimensions (i.e. basic information, family situation, purchase intention, financial situation), we then design an entropy-based weight allocation algorithm to calculate attribute weights, which can explore implicit heterogeneous relationships. Finally, we use the attribute weights, expert knowledge and a representation technology called "bag of attributes" to construct a buyer-specific feature representation and develop a real estate buyer profiling system (REBS).

Our major contributions are described as follows:

- We propose a multi-attribute decision making approach for real estate buyer profiling. To the best of our knowledge, it is the first data-driven user profile work applied in the real estate domain.
- We propose an enriched Bi-LSTM conditional random field (EB-CRF) model, where part-of-speech (POS) tags and named entities are considered to extract key phrases in the sentence more accurately.
- We design an entropy-based weight allocation algorithm to construct attribute-awareness buyer feature representations, which can help explore implicit heterogeneous relationships in structured data.
- We develop a real estate user buyer profiling system (REBS) and demonstrate the excellence of our approach through empirical analysis.

2 Related work

2.1 Multi-attribute decision making

Decision making has broad application prospects in various domains and attracts many scholars to research. Zhang et al. [42] utilized decision tree ensemble classifiers, where appropriate features from each decision tree are selected to participate in the decision-making process. Since the best decision is usually associated with several attributes and each attribute contributes differently, multi-attribute decision making (MADM) aims to find the best decision by assigning different weights to attributes. In other words, attribute weighting is the most important subtask in MADM.

Researches on attribute weighting are usually divided into three categories: subjective, objective and hybrid [8, 35]. Subjective methods depend largely on the expert preference for attributes, including the methods of linear programming, mathematical programming. Horowitz et al. [19] proposed a linear programming (LP) model as an alternative to assess periodic performance appraisal (PA) of subordinates. Deng et al. [11] designed a mathematical programming model that adopt pairwise alternative comparison. Objective methods use an objective decision matrix to determine attribute weights, including the fuzzy entropy method [6, 7], standard deviation (SD) method [10, 35]. However, subjective methods contain abundant subjective expert information that is hard to evaluate, while objective methods may be ineffective due to a lack of expert knowledge. Therefore, hybrid methods are proposed to solve the above issues, which uses expert preference and objective decision matrix to jointly produce attribute weights. Ma et al. [28] formed a two-objective programming model with both subjective and objective information. Fan et al. [15] integrated the fuzzy preference on alternatives and the objective matrix into a general framework, while Wang et al. [36] further considered multiplicative preference relations of the decision-maker on attribute weights.

In this paper, we design a hybrid method to combine the advantages of both subjective and objective methods, where expert knowledge and entropy methods are adopted to realize the MADM approach.

2.2 User Profiling

With the rapid development of the Internet, numerous user information has emerged and user profiling has become a popular data analysis method. User profiles can understand the needs of users through describing users' overall behavior features, and further carry out precise personalized marketing to potential targets. Therefore, user profiling has a wide range of application prospects. Mezghani et al. [29] summarized the characteristics of the social user and tag-based profile modeling and updating techniques. Constantinides et al. [9] used the interaction logs between users and news apps to generate user profiles. Based on computational linguistic features, Hu et al. [16] proposed a psychological modeling method to explore the potential relationship between users' social behaviors on Sina Weibo. Sun et al. [33] proposed an early-warning framework based on clustering methods and association rule mining for online learners. Wu et al. [37] applied user profiles to the field of financial accounting, where expenditure data was used to fill in revenue data. Wu et al. [38] generated user profiles for developers' programming behaviors. Diao et al. [13] studied user profiles from the perspective of transfer learning, where knowledge can be transferred from one social network to another. Kong et al. [23] profiled scientific researchers from five perspectives (i.e. article-centered, author-centered, venue-centered,

institution-centered factors and temporal factors), explaining how to identify and evaluate key factors to improve scientific influence. Cai et al. [2] profiled social media users from the perspective of social networks, which involved three important factors: social connection, spatial connection, and preference-based similarity connection. Kong et al. [24] profiled university based on academic graph and proposed a deep representation clustering model. Li et al. [26] profiled the social media community from the perspective of social influence, which can help enterprises to promote potential customers with different types. In this paper, we applied user profiles to the real estate domain, which can improve buyers' satisfaction and reduce the cost of enterprise promotion.

2.3 Information extraction

With the rapid development of the Internet, the information generated by users is increasing exponentially, which leads to a challenging issue: how to filter out valuable information from massive data.

Information extraction (IE) is the task of extracting information from unstructured data, where key phrase extraction (KPE) is an important subtask. KPE aims to extract key phrases (e.g. entities, relations, events) from natural language text. Hasan et al. [18] introduced the key phrase extraction task and its basic process. Zhang et al. [41] extracted key phases with Bi-RNN, which combined keywords and context information. Chen et al. proposed an encoder-decoder framework to capture deep semantics of content [3], and two additional key phrase constraints are further introduced [4].

In addition, representation learning (RL) is an important component as it enables the model to use deep learning methods. RL aims to find a better way to represent data, which is a numerical vector in the predefined vector space. "Bag of words" is a representation learning approach to represent a sentence as a vector, where the dimension is the size of the vocabulary, and the word's corresponding position refers to its occurrence times in the sentence. However, the method above can't express the semantic relevance between words and may cause dimension disaster. Therefore, distributed representation (e.g. GloVe [31], fastText [22], BERT [12]) is proposed to address the problem. Some representation learning methods have been proposed to explore the potential relations between elements in the task. Du et al. [14] modeled highly unstructured user-generated content based on character embedding and attention-based neural networks. Song et al. [32] deep reconstructed the original data of "exercise-to-concept" into "exercise-to-exercise" and "concept-to-concept", and its potential interactive influences were learned through the process of inference and generation. Both Li et al. [27] and Chen et al. [5] proposed a novel network representation learning approach to effectively capture highly nonlinear network topological structure and attribute information. Hou et al. [20] also provided a systematic overview of network embedding techniques. Furthermore, there are still some researches constructing the representation of objects in heterogeneous networks with graph convolutional network (GCN) [40], topology adapted smoothing [5] and dynamic evolving graphs [39].

In this paper, we perform the information extraction task on unstructured data to obtain structured data for subsequent user profiling.

3 Methodology

In this section, we introduce the overall framework of our proposed approach. We first give the task definition and then describe the architecture of our approach in detail.

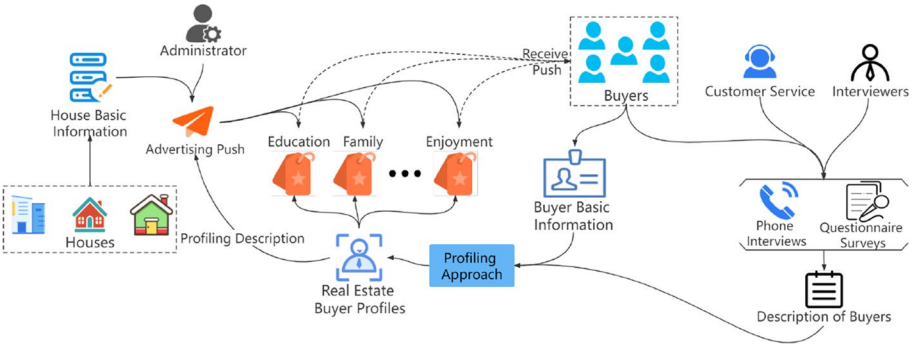


Fig. 1 Task of our real estate buyer profiling

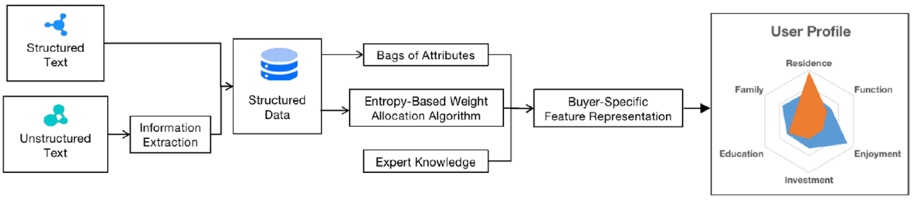


Fig. 2 Overview of our multi-attribute decision making approach

3.1 Task Definition

As shown in Figure 1, we receive information from different sources (e.g. questionnaire surveys, telephone interviews), real estate buyer profiling aims to abstract user behavioral characteristics through analyzing massive data and advertising suitable houses to target customer groups.

In this paper, we treat the task of real estate buyer profiling as a multi-attribute decision making (MADM) problem, which can be defined as a triple (A, C, W). Specifically, A is a set of pre-defined real estate buyer labels, $C = \{c_1, c_2, ..., c_t\}$ is a set of attributes, $W = \{w_1, w_2, ..., w_t\}$ is a set of attribute weight vectors, where each attribute weight in $w_t = \{w_{t1}, w_{t2}, ..., w_{tk}\}$ is non-negative and $\sum_{j=1}^{k} w_{tj} = 1$. The goal of the MADM task is to find the best attribute weight vector for attributes to better abstract the user's behavioral characteristics.

3.2 Approach Architecture

Figure 2 shows the architecture of our proposed MADM approach, which will be introduced in this section generally. Firstly, we separate the unstructured data and perform the information extraction (IE) task on it to obtain structured data, where original structured data can be refined in this step. Then, we apply the "bag of attributes", a representation technology similar to "bag of words", to these reorganized structured data to obtain the basic buyer feature representations. With the help of expert knowledge, we design an entropy-based weight allocation algorithm to obtain attribute weights for each

real estate buyer. Finally, we construct a buyer-specific feature representation for real estate buyer profiling, which can be visualized by a radar chart.

3.3 Information Extraction

Information extraction (IE) task aims to extract structured data from unstructured data, which is a crucial component of our MADM approach. The details of IE task are shown in Figure 3.

Given a descriptive sentence, we first obtain the sentence S with n words $S = \{s_1, s_2, ..., s_n\}$ and its corresponding entity or part of speech (POS) tags $T = \{t_1, t_2, ..., t_n\}$ with the help of lexical analysis of Chinese (LAC) [21]. Then, we feed it into our proposed EB-CRF model to extract key phrases. Specifically, we obtain the enriched hidden representations X by concatenating word embedding R and tag representations E, which are generated through fastText [22]:

$$R = fastText(S) \tag{1}$$

$$E = fastText(T) \tag{2}$$

$$X = [E : R] \tag{3}$$

where [:] is a concatenation operator. We use BiLSTM to encode contextual semantic information, and further add some implicit constraint rules with conditional random fields (CRF) [25]:

$$H = BiLSTM(X) \tag{4}$$

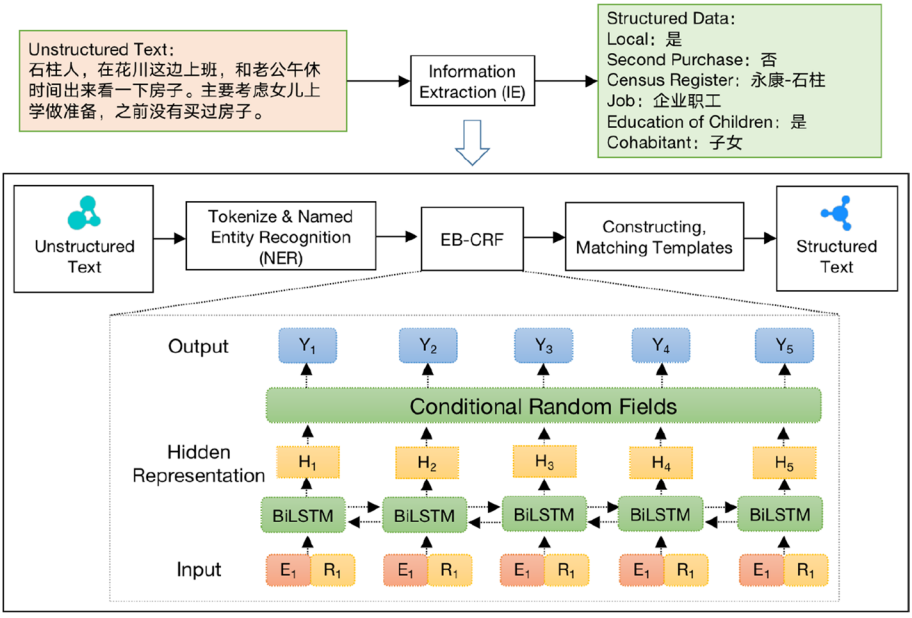


Fig. 3 Details of information extraction task

$$H' = CRF(H) \tag{5}$$

Next, we regard the KPE task as a sequence tagging problem and predict key phrases in the sentence with the help of predictive label set $L = \{B, I, O, E, S, U\}$:

$$Y = \sigma(W_k H + b_k) \tag{6}$$

where σ is an activation function (e.g. Tanh, Sigmoid, ReLu), W_k and b_k are the parameters to be trained. In addition, elements B, I, O, E, S, U in L represent key phrase's begin, key phrase's inside, key phrase's outside, key phrase's end, single word key phrase, the useless word in the key phrase, respectively.

After obtaining key phrases in the sentence, we analyze the corpus and construct a series of templates based on regular expressions and fuzzy theory. Then, we match the extracted key phrases with these templates to obtain the structured data. Specifically, for each data field, we construct a correlation matrix Q, where Q_{ij} represents the score obtained when the regular expression j matches the phrase for the optional value i. We then calculate the score for each optional value and take the value with highest score as the final matching result:

$$F = M * Q^T \tag{7}$$

where M is a matching vector consisting of 1 (matched) and 0 (mismatched). Next, we take the original structured data as the main body and supplement it with the structured data obtained by the IE task. Finally, we organize these structured data from different sources to form a structured dataset.

3.4 Representation learning

After obtaining reorganized structured data, we apply a representation learning approach to construct buyer-specific feature representations for real estate buyer profiling.

Firstly, we classify the attributes into four general categories: basic information, family situation, purchase intention, financial situation, which are shown in Figure 4. Specifically, basic information refers to the buyer's natural and social attributes; family situation aims to understand the family of real estate buyers and their potential demands; purchase intention refers to the buyer's preference for the house; financial situation aims to understand buyers' affordability to purchase.

Then, based on the four general categories above, we can obtain the basic buyer feature representations B and attribute-awareness buyer feature representations U with the representation technology "bag of attributes" and GloVe [31], respectively. Specifically, we construct category-specific buyer feature representations and attribute-awareness buyer feature representations for each general category and then concatenate them. It is worth noting that "bag of attributes" is a technique that uses an attribute vocabulary to convert a buyer's attributes into a sparse numerical vector, which is similar to "bag of words" but the concept of the word is replaced by attribute.

Next, we design an entropy-based weight allocation algorithm to obtain attribute weights, which are shown in Algorithm 1. As line 2 to line 17 of Algorithm 1 shows, we first cluster user representations with k-means algorithm to obtain K groups of real estate users with similar features. Then, we introduce the concept of information entropy in line 18, which is a quantitative measure of information. The more complex the object,

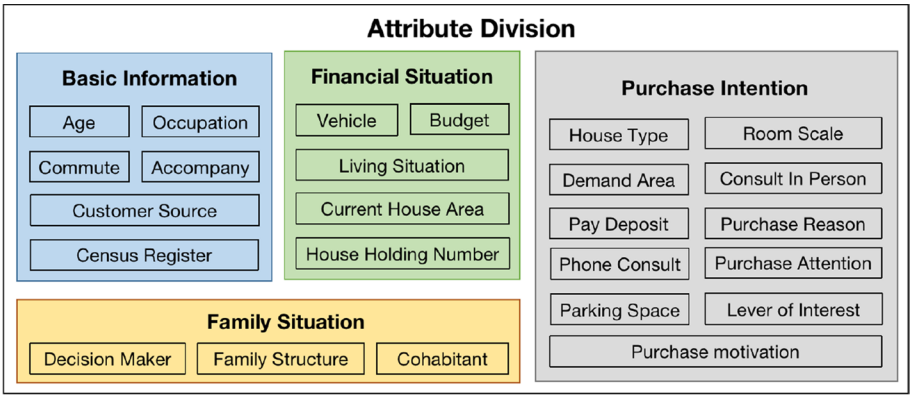


Fig. 4 Attribute division of real estate buyers

the greater the information entropy, and vice versa. Based on the above theory, we assign weight to each attribute and influence the proportion of weight by integrating expert knowledge, as shown in line 19 to line 26.

Algorithm 1 Entropy-based weight allocation algorithm for real estate buyer representation construction.

input:
Expert knowledge EK, attribute set C and attribute-awareness buyer feature representations U, where U_i donates a buyer representation for real estate buyer i.
output:
Attribute weight set W, where $W_i = \{w_{i1}, w_{i2}, ..., w_{ik}\}$ donates a non-negative attribute weight vector for buyer i and $\sum_{j=1}^{k} w_{ij} = 1$.

```
1: Randomly selected K points as cluster centers
2: while K cluster centers changed or below max Iteration do
3:     minDistance, belongCluster = float("inf"), -1
4:     for U_i in U do
5:         for k in K do
6:             distance = calculate(U_i, k)
7:             if distance < minDistance then
8:                 minDistance = distance
9:                 belongCluster = k
10:            end if
11:        end for
12:        Assign U_i to belongCluster
13:    end for
14:    for k in K do
15:        Recalculate and update cluster center
16:    end for
17: end while
18: E(X) = −∑_{x∈X} p(x)logp(x)
19: for k in K do
20:    for j in C do
21:        for user i in cluster k do
22:            W_ij = E(∑_{t=1}^{K} t_j) / E(k_j)
23:        end for
24:    end for
25:    W_i = EK * Normalize(W_i)
26: end for
```

Finally, we weighted sum the attribute weights and basic buyer feature representations to obtain the final buyer-specific feature representations. According to final buyer-specific feature representations, we choose the weight that is greater than the threshold as the characteristic label and visualize it by radar chart. It is worth noting that a real estate buyer may have multiple characteristic labels.

4 Experiment

4.1 Datasets and metrics

We employ real-world real estate sales dataset to demonstrate the effectiveness of our MADM approach. The dataset we used in our experiments is derived from four different real estate of a company in China, and each dataset contains 500 data and different scale data fields. Due to various reception methods and non-standard questionnaire filling, the format and fields of the dataset are inconsistent, we preprocess the original dataset, where we retain important data fields and ensure its consistency. Table 1 shows the pre-processed data of the original dataset, which can be roughly divided into structured data and unstructured data (i.e. description of buyers).

There is also a large number of vacancies in the dataset. For example, the vacancy rate of the "Census Register" field is 21.7%, the "Demand Area" field is 25.05%, the "Occupation" field is 25.10% and the "Cohabitant" field is up to 55.90%, which can affect the performance of our approach. Meanwhile, we notice that the description of buyers in the unstructured data may provide additional information to fill in vacancies values or further expand the data filed. Besides, due to the individual differences of interviewers, there is no unified format for describing customers in different styles. Some descriptions are the interviewer's oral paraphrase that contains a large number of meaningless modal particles. Therefore, we apply our EB-CRF model for the KPE task to extract key phrases in the sentence and we obtain a reorganized dataset with 27 data fields and 238 attributes for following real estate buyer profiling through sorting and induction.

We indirectly prove the effectiveness of our proposed MADM approach by evaluating the performance of the KPE task. To evaluate the performance of our proposed KPE approach, we use precision (P), recall (R), and F1-score (F1) as the metrics:

$$R = \frac{TP}{TP + FN} \tag{8}$$

$$P = \frac{TP}{TP + FP} \tag{9}$$

$$F1 = \frac{2PR}{P + R} \tag{10}$$

where TP refers to the number of positive cases correctly predicted, FP refers to the number of negative cases incorrectly predicted, TN refers to the number of negative cases correctly predicted, and FN refers to the number of positive cases incorrectly predicted.

Table 1 The pre-processed data of the original dataset

Structured Data			
Data Field	Name	Purchase Attention	Phone Consult
	Sex	Room Scale	Occupation
	Age	House Type	Decision Maker
	Level of Interest	Purchase Motivation	Vehicle
	Census Register	Family Structure	Visit Type
	Living Area	Pay Deposit	Customer Source
	Working Area	Consult In Person	Living Situation
	Demand Area	Parking Space	Purchase Reason
	Budget	Current House Area	Cohabitant
Unstructured Data			
Descriptive Text	Description of buyers from telephone interviews and questionnaire surveys.		

4.2 Experimental Settings

We tune the hyper-parameters through a large number of experiments, and the relevant results are shown in Figure 5, where the x-coordinate is the value of the hyper-parameter and the y-coordinate is the f1 score. As shown in Figure 5(a) and Figure 5(b), since the dimension of NER and hidden dimension achieve the best F1 score at 24 and 100, we set them to 24 and 100. Since SGD is efficient and convergent fast, we choose it as the optimizer and the learning rate is 0.016, which can be observed in Figure 5(c). For the robustness of our model, we set the dropout rate to 0.1. We also set the threshold to 0.7. We select the best model according to the best F1 score.

4.3 Baselines

To demonstrate the effectiveness of EB-CRF for the KPE task, we compare it with the following baselines. The hyper-parameters for baselines are set to the optimal values as reported in their papers.

- Joint-layer RNN [41] is an end-to-end model, which jointly processes the keyword ranking and keyphrase generation task.
- LSTM-LSTM [43] is an end-to-end model for joint extraction of entities and their relations. It contains two Bi-LSTM layers for encoding and decoding the sentence, respectively. In addition, biased objective function is adopted to enhance the association between related entities.
- BiLSTM-CRF [1] combine the advantage of both CRF [25] and BiLSTM for keyphrase extraction. Specifically, CRF can capture label dependencies through a transition parameter matrix and BiLSTM can capture implicit semantics in the sentence through the long-distance dependencies.
- CNN-BiLSTM-CRF is an end-to-end model, which contains a CNN layer, a BiLSTM layer and a CRF [25] layer for key phrase extraction.

– MHA-BiLSTM-CRF is an end-to-end model, which contains multi-head attention mechanism (MHA) [34], a BiLSTM layer and a CRF [25] layer for key phrase extraction.

4.4 Results of key phrase extraction

Table 2 compares EB-CRF with the state-of-the-art approaches for the KPE task. We can observe that EB-CRF outperforms the results of other baseline approaches significantly in recall and F1-score. This result indicates that our model can capture the deep semantic features in the sentence and better locate crucial information, which is achieved by considering POS tags and named entities as a part of the input in our model.

4.5 Analysis

4.5.1 Ablation study

To investigate the effects of different components in EB-CRF, we conduct an ablation study for the KPE task. The experiment results are shown in Table 3. After removing the component of BiLSTM or CRF, the performance of EB-CRF is reduced sharply, which indicates that BiLSTM can capture long-distance dependent semantic information in sentences and CRF can learn constraints between words. Furthermore, after removing the embedding of POS tags and name entities, we can observe that the performance of EB-CRF is reduced apparently in recall and F1-score. We infer that POS tags and name entities can help understand the semantic of the words in the sentences and find the key information more comprehensively, which contributes little to predicting the results accurately.

4.5.2 Comparison of Single-word and Multi-word Key Phrase

We compare the performance of EB-CRF with the previous model BiLSTM-CRF for the following two settings in Table 4: Single-Word: Key phrase is a single word span, Multi-Word: Key phrase is a multi-word span. For the multi-word setting, our method shows consistent improvement in terms of both precision and recall score, which results in the improvement of the F1 score. When we compare the evaluations for single-word key phrases, our model achieves more significant improvements for F1 scores. Compared to precision, our recall shows greater improvement over the BiLSTM-CRF approach. This result indicates that our approach can better extract single-word key phrases in sentences, owing to the embedding of POS tags and name entities.

4.5.3 Results of information padding

We experiment on the result of information padding. The results are shown in Figure 6. We can observe that the results performed differently in the different data fields. Generally speaking, our method uses unstructured data to fill in missing values in structured data, and its accuracy can achieve more than 54%, which brings a certain degree of supplement to the original default data. It's worth noting that due to the arbitrariness of natural language

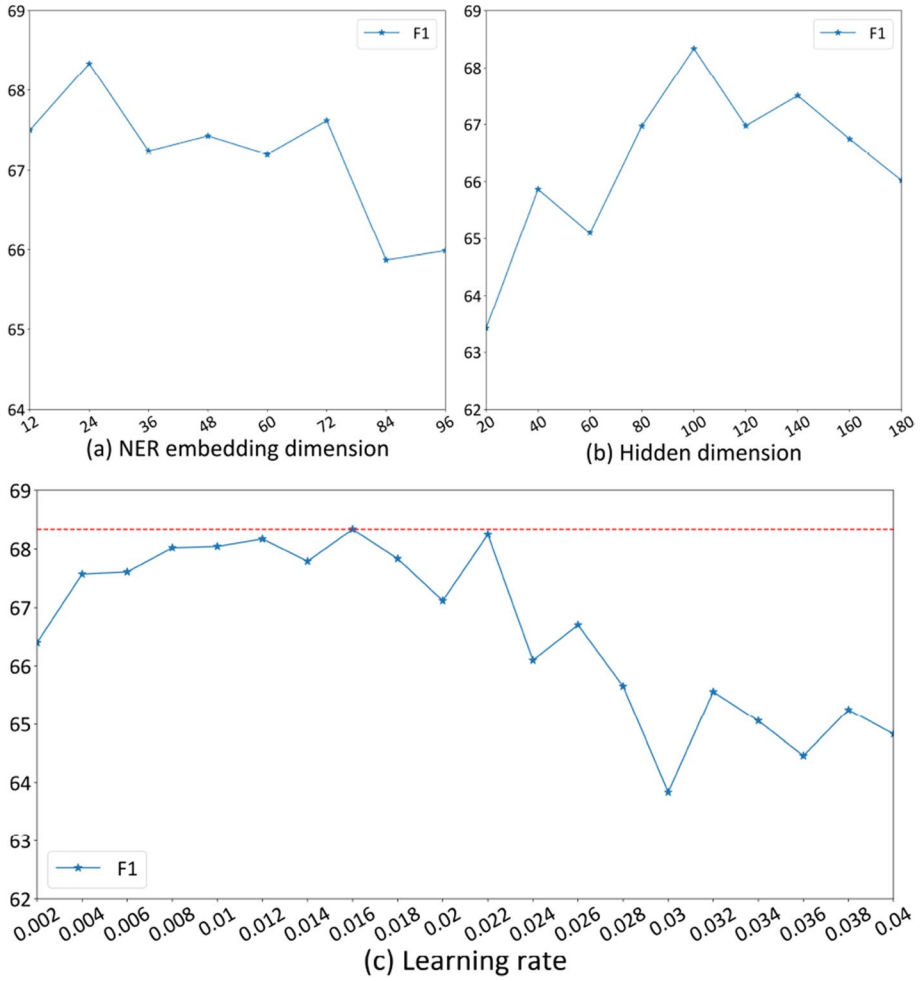


Fig. 5 Experiment of parameter setting

expression, many omissions (e.g. subject, verb, object) and ambiguities (e.g. maybe, perhaps) may occur, resulting in poor performance of capturing key information.

4.5.4 Analysis of real estate buyer groups

To illustrate the effectiveness of our proposed MADM approach, we analyze the frequency results of attributes in the target real estate buyer groups, which are shown in Figure 7. In practice, we profile the characteristics of real estate buyers from six perspectives (i.e. family, education, investment, enjoyment, function and residence) and set the threshold to 0.7.

As shown in Figure 7(a), native people tend to improve their quality of life from the perspective of investment, enjoyment, function and education, while outsiders

tend to settle down and set up families. Figure 7(c) shows the age distribution of the interviewed real estate buyers. On the whole, the age distribution is uniform, and real estate buyers for function, investment and education purposes are relatively older. Figure 7(h) shows the occupation distribution of the interviewed real estate buyers. Specifically, enterprise employees usually have the characteristics of residence, teachers pay more attention to education and freelancer usually have no characteristics of the family. Figure 7(b) and (d) show the details of family situation. Generally speaking, the common family number of families is 3 and the most common cohabitant is children and spouse, while real estate buyers with family characteristics tend to have more than 3 members and pay more attention to satisfying family needs. Figure 7(e) and Figure 7(i) indicate that real estate buyers with family or residence characteristics tend to have fewer financial reserves and purchase for self-living, while real estate buyers with enjoyment, investment, function and education characteristics tend to have adequate financial reserves and purchase for improving quality of life. We also analyzed the motivations and attentions of real estate buyers in Figure 7(f), Figure 7(g) and (j). Specifically, real estate buyers with family or residential characteristics have never bought a house before, so they attach great importance to price; real estate buyers with enjoyment and function characteristics usually own a house, but poor supporting facilities may lead them to intension to buy a new house with sufficient supporting facilities; real estate buyers with investment characteristics usually have multiple assets, therefore, they pay more attention to the balance of overall factors; real estate buyers with education characteristics usually pay more attention to children's education, therefore, their main appeal is school district housing.

4.5.5 Case study

A real estate buyer profiling case is given in Figure 8, where Figure 8(a) shows the overall real estate buyer profile and Figure 8(b)(c) shows the constitution of its most outstanding characteristics: education and function. As we described earlier, we set the threshold to 0.7 and the attribute exceeding the threshold will be regarded as the user's characteristic label. Therefore, Figure 8 shows the profile of a real estate buyer with education characteristics. Further, we analyze the real estate buyer with the two most outstanding characteristics. It is worth noting that the charts mainly show the top 7 factors with the highest correlation and others are a collection of lower-related factors. Real estate buyer with education

Table 2 The experiment results on the KPE task (%)

Model	P	R	F1
Joint-layer RNN	53.57	43.74	48.15
LSTM-LSTM	59.63	59.50	59.76
CNN-BiLSTM-CRF	68.37	61.76	64.90
MHA-BiLSTM-CRF	67.06	63.08	65.00
BiLSTM-CRF	**69.38**	62.75	65.90
EB-CRF	69.22	**67.48**	**68.34**

We highlight the best results in bold

Table 3 Ablation test on the KPE task (%)

Model	P	R	F1
EB-CRF	69.22	**67.48**	**68.34**
w/o CRF	59.63	57.14	58.36
w/o BiLSTM	52.73	52.09	52.40
w/o POS&NER	**69.38**	62.75	65.90

We highlight the best results in bold

Table 4 Analysis with different modes on the EB-CRF model (%)

Model	Multi-Word			Single-Word		
	P	R	F1	P	R	F1
Joint-layer-RNN	49.19	41.47	45.00	75.20	53.11	62.25
LSTM-LSTM	56.89	57.98	57.43	71.86	67.80	69.77
CNN-BiLSTM-CRF	67.64	60.16	63.68	71.18	68.36	69.74
MHA-BILSTM-CRF	65.43	62.48	63.92	74.36	65.54	69.67
BiLSTM-CRF	**68.29**	63.17	65.63	74.48	61.02	67.08
EB_CRF	67.03	**66.58**	**66.80**	**79.25**	**71.19**	**75.00**

We highlight the best results in bold

characteristics attaches great importance to education, are over 40 years old, live with their children and prefer school district houses. Real estate buyers with function characteristics usually own a house and have a good budget, so they pay more attention to the type of house.

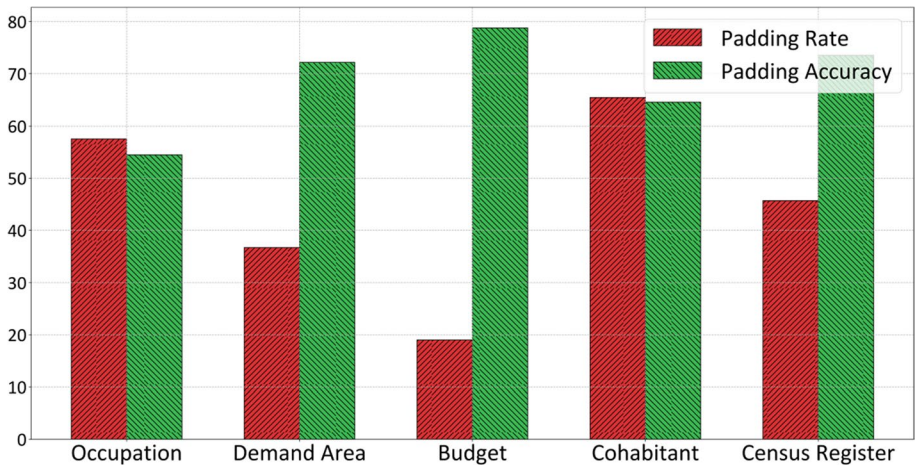
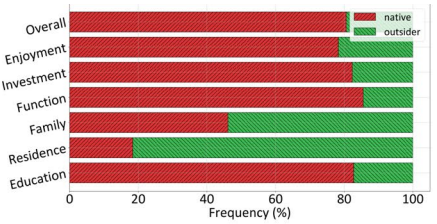
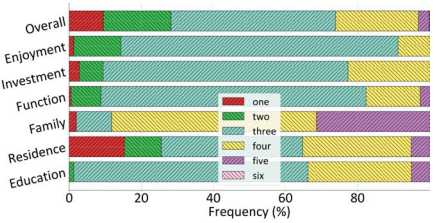


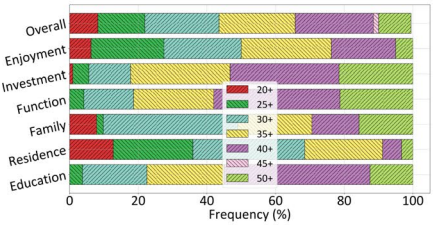
Fig. 6 Results of information padding (%)



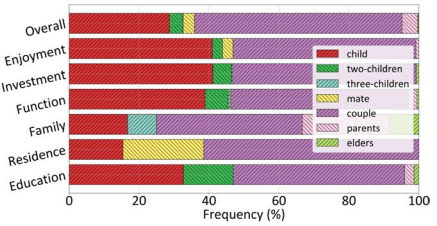
(a) Census Register



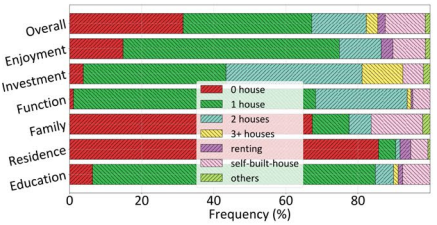
(b) Family Structure



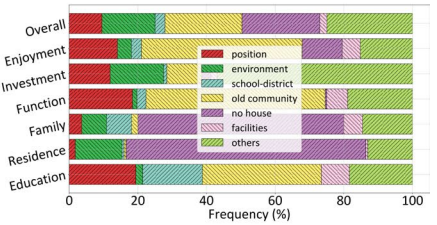
(c) Age



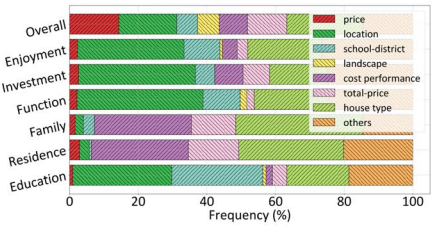
(d) Cohabitant



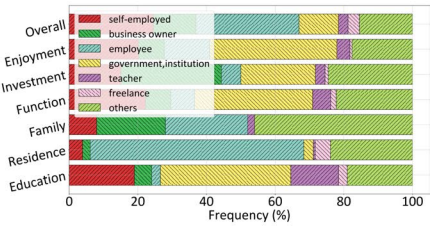
(e) Living Situation



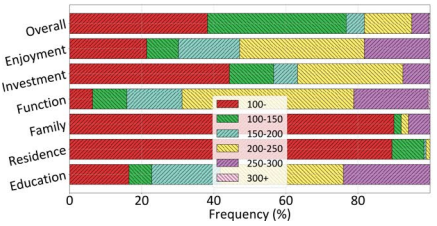
(f) Purchase Reason



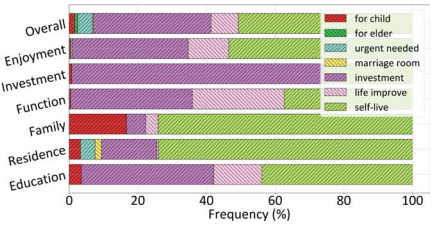
(g) Purchase Attention



(h) Occupation



(i) Budget



(j) Purchase Motivation

Fig. 7 Analysis of Real Estate Buyer Groups

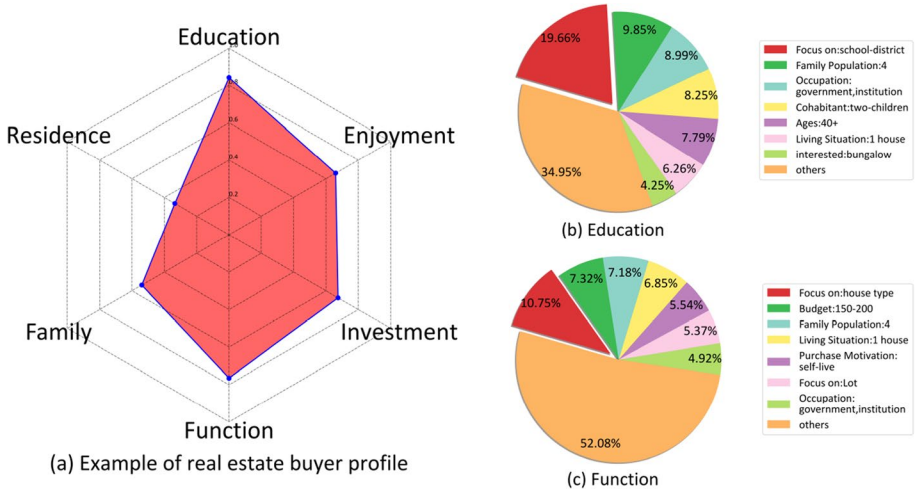


Fig. 8 Case study of real estate buyer with educational characteristics

5 Conclusion

In this paper, we propose a novel multi-attribute decision making (MADM) approach for real estate buyer profiling, which can help enterprises accurately locate target customer groups. Firstly, we reorganize the dataset with our proposed enriched bi-directional long short-term memory conditional random field (EB-CRF) model and man-made templates. Based on four general dimensions, we then use "bag of attributes" and an entropy-based weight allocation algorithm to obtain the buyer-specific feature representations. Finally, we obtain the real estate buyer profiles and develop a real estate buyer profiling system (REBS). Extensive experimental results indicate that our model outperforms strong baselines significantly and achieves state-of-the-art performance. Future work includes performing finer-grained user profiling methods and applying novel natural language methods to better deal with informal text on social occasions.

Acknowledgements We would like to thank the anonymous reviewers for their insightful comments. This work was supported by the National Natural Science Foundation of China (No. 62176234, 62072409, 62176078, 61701443)

Declarations

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Alzaidy, R., Caragea, C., Giles, C.L.: Bi-lstm-crf sequence labeling for keyphrase extraction from scholarly documents. In: Liu L., White R.W., Mantrach A., Silvestri F., McAuley J.J., Baeza-Yates R., Zia L. (eds.) The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019, pp. 2551–2557. ACM (2019). https://doi.org/10.1145/3308558.3313642

2. CAI, T., Li, J., Mian, A.S., li, R., Sellis, T., Yu, J.X.: Target-aware holistic influence maximization in spatial social networks. IEEE Transactions on Knowledge and Data Engineering, 1–1 (2020). https://doi.org/10.1109/TKDE.2020.3003047
3. Chen, W., Chan, H.P., Li, P., King, I.: Exclusive hierarchical decoding for deep keyphrase generation. In: Jurafsky D., Chai J., Schluter N., Tetreault J.R. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pp. 1095–1105. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.103
4. Chen, J., Zhang, X., Wu, Y., Yan, Z., Li, Z.: Keyphrase generation with correlation constraints. In: Riloff E., Chiang D., Hockenmaier J., Tsujii J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, pp. 4057–4066. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/d18-1439
5. Chen, J., Zhong, M., Li, J., Wang, D., Qian, T., Tu, H.: Effective deep attributed network representation learning with topology adapted smoothing. IEEE Transactions on Cybernetics, 1–12 (2021). https://doi.org/10.1109/TCYB.2021.3064092
6. Chen, T., Li, C.: Determining objective weights with intuitionistic fuzzy entropy measures: A comparative analysis. Inf. Sci. **180**(21), 4207–4222 (2010). https://doi.org/10.1016/j.ins.2010.07.009
7. Chen, T., Li, C.: Objective weights with intuitionistic fuzzy entropy measures and computational experiment analysis. Appl. Soft Comput. **11**(8), 5411–5423 (2011). https://doi.org/10.1016/j.asoc.2011.05.018
8. Chin, K., Fu, C., Wang, Y.: A method of determining attribute weights in evidential reasoning approach based on incompatibility among attributes. Comput. Ind. Eng. **87**, 150–162 (2015). https://doi.org/10.1016/j.cie.2015.04.016
9. Constantinides, M., Dowell, J.: A framework for interaction-driven user modeling of mobile news reading behaviour. In: Mitrovic T., Zhang J., Chen L., Chin D. (eds.) Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization, UMAP 2018, Singapore, July 08-11, 2018, pp. 33–41 (2018). https://doi.org/10.1145/3209219.3209229
10. Deng, H., Yeh, C., Willis, R.J.: Inter-company comparison using modified TOPSIS with objective weights. Comput. Oper. Res. **27**(10), 963–973 (2000). https://doi.org/10.1016/S0305-0548(99)00069-6
11. Deng, M., Xu, W., Yang, J.: Estimating the attribute weights through evidential reasoning and mathematical programming. Int. J. Inf. Technol. Decis. Mak. **3**(3), 419–428 (2004). https://doi.org/10.1142/S0219622004001124
12. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J., Doran C., Solorio T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/n19-1423
13. Diao, M., Zhang, Z., Su, S., Gao, S., Cao, H.: UPON: user profile transferring across networks. In: d'Aquin M., Dietze S., Hauff C., Curry E., Cudré-Mauroux P. (eds.) CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020, pp. 265–274 (2020). https://doi.org/10.1145/3340531.3411964
14. Du, J., Michalska, S., Subramani, S., Wang, H., Zhang, Y.: Neural attention with character embeddings for hay fever detection from twitter. Health Inf. Sci. Syst. **7**(1), 21 (2019). https://doi.org/10.1007/s13755-019-0084-2
15. Fan, Z., Ma, J., Zhang, Q.: An approach to multiple attribute decision making based on fuzzy preference information on alternatives. Fuzzy Sets Syst. **131**(1), 101–106 (2002). https://doi.org/10.1016/S0165-0114(01)00258-5
16. Gu, H., Wang, J., Wang, Z., Zhuang, B., Su, F.: Modeling of user portrait through social media. In: 2018 IEEEx International Conference on Multimedia and Expo, ICME 2018, San Diego, CA, USA, July 23-27, 2018, pp. 1–6 (2018). https://doi.org/10.1109/ICME.2018.8486595
17. Han, Y., Zhang, H., Zhao, Y.: Structural evolution of real estate industry in china: 2002–2017. Structural Change and Economic Dynamics **57**, 45–56 (2021). https://doi.org/10.1016/j.strueco.2021.01.010
18. Hasan, K.S., Ng, V.: Automatic keyphrase extraction: A survey of the state of the art. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers, pp. 1262–1273. The Association for Computer Linguistics (2014). https://doi.org/10.3115/v1/p14-1119
19. Horowitz, I., Zappe, C.: The linear programming alternative to policy capturing for eliciting criteria weights in the performance appraisal process. Omega **23**(6), 667–676 (1995). https://doi.org/10.1016/0305-0483(95)00039-9

20. Hou, M., Ren, J., Zhang, D., Kong, X., Zhang, D., Xia, F.: Network embedding: Taxonomies, frameworks and applications. Computer Science Review **38**, 100,296 (2020). https://doi.org/10.1016/j.cosrev.2020.100296
21. Jiao, Z., Sun, S., Sun, K.: Chinese lexical analysis with deep bi-gru-crf network. arXiv:1807.01882 (2018)
22. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: Lapata, M., Blunsom, P., Koller A. (eds.) Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers, pp. 427–431. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/e17-2068
23. Kong, X., Zhang, J., Zhang, D., Bu, Y., Xia, F.: The gene of scientific success. ACM Trans. Knowl. Discov. Data **14**(4), 41:1-41:19 (2020). https://doi.org/10.1145/3385530
24. Kong, X., Li, J., Wang, L., Shen, G., Sun, Y., Lee, I.: Recurrent-dc: A deep representation clustering model for university profiling based on academic graph. Future Generation Computer Systems **116**, 156–167 (2021). https://doi.org/10.1016/j.future.2020.10.019
25. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Brodley C.E., Danyluk A.P. (eds.) Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001, pp. 282–289. Morgan Kaufmann (2001)
26. Li, J., Cai, T., Deng, K., Wang, X., Sellis, T., Xia, F.: Community-diversified influence maximization in social networks. Information Systems **92**, 101,522 (2020). https://doi.org/10.1016/j.is.2020.101522
27. Li, Z., Wang, X., Li, J., Zhang, Q.: Deep attributed network representation learning of complex coupling and interaction. Knowl. Based Syst. **212**, 106,618 (2021). https://doi.org/10.1016/j.knosys.2020.106618
28. Ma, J., Fan, Z., Huang, L.: A subjective and objective integrated approach to determine attribute weights. Eur. J. Oper. Res. **112**(2), 397–404 (1999). https://doi.org/10.1016/S0377-2217(98)00141-6
29. Mezghani, M., Zayani, C.A., Amous, I., Gargouri, F.: A user profile modelling using social annotations: a survey. In: Mille A, Gandon F, Misselis J, Rabinovich M, Staab S. (eds.) Proceedings of the 21st World Wide Web Conference, WWW 2012, Lyon, France, April 16-20, 2012 (Companion Volume), pp. 969–976 (2012). https://doi.org/10.1145/2187980.2188230
30. Mi, X., Tian, Y., Kang, B.: A hybrid multi-criteria decision making approach for assessing healthcare waste management technologies based on soft likelihood function and d-numbers. Appl. Intell. **51**(10), 6708–6727 (2021). https://doi.org/10.1007/s10489-020-02148-7
31. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Moschitti A., Pang B., Daelemans W. (eds.) Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pp. 1532–1543 (2014). https://doi.org/10.3115/v1/d14-1162
32. Song, X., Li, J., Tang, Y., Zhao, T., Chen, Y., Guan, Z.: Jkt: A joint graph convolutional network based deep knowledge tracing. Information Sciences **580**, 510–523 (2021). https://doi.org/10.1016/j.ins.2021.08.100
33. Sun, Y., Chai, R.: An early-warning model for online learners based on user portrait. Ingénierie des Systèmes d Inf. **25**(4), 535–541 (2020). https://doi.org/10.18280/isi.250418
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon I., von Luxburg U., Bengio S., Wallach H.M., Fergus R., Vishwanathan S.V.N., Garnett R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pp. 5998–6008 (2017). https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
35. Wang, Y., Luo, Y.: Integration of correlations with standard deviations for determining attribute weights in multiple attribute decision making. Math. Comput. Model. **51**(1–2), 1–12 (2010). https://doi.org/10.1016/j.mcm.2009.07.016
36. Wang, Y., Parkan, C.: A general multiple attribute decision-making approach for integrating subjective preferences and objective information. Fuzzy Sets Syst. **157**(10), 1333–1345 (2006). https://doi.org/10.1016/j.fss.2005.11.017
37. Wu, Y., Wang, R., Dai, W., Dong, S., You, X., You, H., Liu, L.: User portraits and investment planning based on accounting data. In: 2020 IEEE International Conference on Services Computing, SCC 2020, Beijing, China, November 7-11, 2020, pp. 404–411 (2020). https://doi.org/10.1109/SCC49832.2020.00059

38. Wu, Y., Yu, P.: User portrait technology based on stacking mode. In: IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress, DASC/PiCom/CBDCom/CyberSciTech 2020, Calgary, AB, Canada, August 17-22, 2020, pp. 245–250 (2020). https://doi.org/10.1109/DASC-PICom-CBDCom-CyberSciTech49142.2020.00051
39. Xue, G., Zhong, M., Li, J., Chen, J., Zhai, C., Kong, R.: Dynamic network embedding survey. arxiv: abs/2103.15447 (2021)
40. Yang, Y., Guan, Z., Li, J., Zhao, W., Cui, J., Wang, Q.: Interpretable and efficient heterogeneous graph convolutional network. IEEE Transactions on Knowledge and Data Engineering, 1–1 (2021). https://doi.org/10.1109/TKDE.2021.3101356
41. Zhang, Q., Wang, Y., Gong, Y., Huang, X.: Keyphrase extraction using deep recurrent neural networks on twitter. In: Su J., Carreras X., Duh K. (eds.) Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016, pp. 836–845. The Association for Computational Linguistics (2016). https://doi.org/10.18653/v1/d16-1080
42. Zhang, F., Wang, Y., Liu, S., Wang, H.: Decision-based evasion attacks on tree ensemble classifiers. World Wide Web **23**(5), 2957–2977 (2020). https://doi.org/10.1007/s11280-020-00813-y
43. Zheng, S., Wang, F., Bao, H., Hao, Y., Zhou, P., Xu, B.: Joint extraction of entities and relations based on a novel tagging scheme. In: Barzilay R., Kan M. (eds.) Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, pp. 1227–1236. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/P17-1113