Information Sciences 513 (2020) 98–111

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A service recommendation algorithm with the transfer learning based matrix factorization to improve cloud security

Chao Lei[a], Hongjun Dai[a,*], Zhilou Yu[b], Rui Li[c]

[a] School of Software, Shandong University, No. 1500, Shunhua Road, Jinan, PR China
[b] The Chaoyue Digital Electronics Co. Ltd and the Key Laboratory of Special Computer, Shandong Province, Jinan, PR China
[c] The Technology Center, Inspur Inc. Jinan, PR China

A R T I C L E I N F O

Article history:
Received 1 May 2019
Revised 29 September 2019
Accepted 5 October 2019
Available online 12 October 2019

Keywords:
Service recommendation
Transfer learning
Cloud security

A B S T R A C T

Recommendation system (RS) is designed to provide personalized services based on the users' historical data. It has been applied in various fields and is expected to recommend the suitable services for the different kinds of users. Considering the importance of individual privacy, current users gradually tend not to expose personal information. This means RS may face the highly sparse datasets in the fields of cloud security. In general, the accuracy of recommendation will be improved with the growth of individual data, but the cold start problem is exactly in this contradictory phenomenon: this question evolves to produce sufficiently accurate recommendation result under the data scarcity problem. RS has to recommend services for the rarely historical data users and the latent users might drain along with the production of counter effects. To alleviate data scarcity problem in cloud security environment, this work is to introduce similar domain knowledge based on the transfer learning. Besides, the content and location based methods have been proved that these ideas work under this situation. So, this work also employs *latent dirichlet allocation* (LDA) to analysis the service descriptions and explore the relationship between the content and location information. In this framework, the suitable combination of LDA and word2vec models will balance the accuracy and speed which benefit service recommendation particularly. The related experiments demonstrate the effectiveness on the real word dataset. It can be found that the transfer learning based word2vec model shows the potentiality to explore the relationship between topic words, and improve the LDA algorithm from the content relationship. This proves that in both cold start environment and warm start environment, the proposed algorithm is more robust than other model-based state-of-art methods.

© 2019 Elsevier Inc. All rights reserved.

1. Introduction

In recent years, with the rapid development of cloud computing, *software as a service* (SaaS) is widely used as a method of software delivery and licensing through the Internet without the necessary installation of software on personal computers or intelligent terminals [1]. Services can be accessed online via subscription [2], so the *recommendation system* (RS) gets very important with the increased number of web services [3]. But, the descriptions may be imprecise, even be mali-

* Corresponding author.
E-mail addresses: cn_leichao@hotmail.com (C. Lei), dahogn@sdu.edu.cn (H. Dai), yuzhl@chaoyuesd.com (Z. Yu), lirui01@inspur.com (R. Li).

https://doi.org/10.1016/j.ins.2019.10.004
0020-0255/© 2019 Elsevier Inc. All rights reserved.

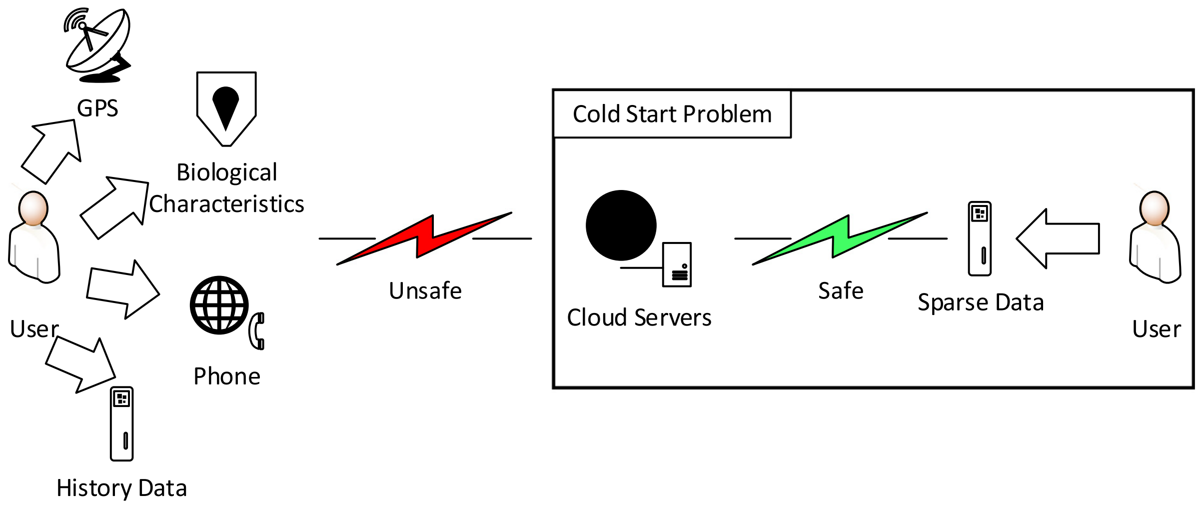

Fig. 1. An example for cold start problem in cloud security. The RS needs to consider the cold start problem in safe situation. The user who frequently explores too much personal data to cloud servers is in danger.

cious probably, then the users should pay more attention to cloud security in SaaS [4]. Furthermore, individual users will still tend not to expose personal information to the Internet. It means few or no individual data will be available for the RS, which leads that the RS may usually face the highly sparse dataset [5]. Then, this brings a typical cold start problem, which is how to design a personalized recommendation without a large amount of users data and let the user be satisfied with the recommendation results [6]. Service commendation has widely employed machine learning algorithms such as *collaborative filtering* (CF) [7]. These methods mainly utilize historical information of user and recommend the services based on the similar preferences. However, the CF may not be robust to deal with the cold start problem that frequently appears in a real world. In addition, these methods also lead to the improper recommendation result due to inaccurate service description considering the content information [8]. So, it is a great challenge due to that the target user's historical information is incomplete. Based on previous efforts, this work presents a machine learning based method to exploit other domain knowledge to alleviate cold start problem in a cloud security scenario, as an extension of the *transfer learning based matrix factorization* (TLMF).

As shown in the Fig. 1, it is unsafe for cloud security to collect personal information multiple times such as GPS information from users. It rise the potential risks of attack, considering with the increase of the interactions between users and cloud servers. Although much effort has been made to ensure the data storage security [9,10], the direct way to improve cloud security is not to expose personal information to provides maximum protection, and the RS needs to consider the fact that few data will be stored in cloud servers. In mathematics, the naive CF based methods may not obtain the historical information. Then the matrix to be decomposed in CF is very sparse and then these methods make the unreliable recommendation. It is not surprised that the naive CF based methods may give the unreliable result for users. So, the key point of the cold start problem is to solve the data scarcity problem.

Many efforts have been taken to alleviate the data scarcity problem in a cloud environment. Lo et al. [11] introduces the location information as the regularization terms for *matrix factorization* (MF) to improve the *quality of service* (QoS), but it only works for the target user not item. Lee et al. [12] develops location-based MF via *preference propagation* (PP), the main idea is to compute similarity via preference propagation and apply MF based on the result of PP and location information. Xie et al. [13] proposals the *time-aware collaborative domain regression* (TCDR) based the idea of the *collaborative topic regression* (CTR). Yuan et al. [14], Rupasingha and Paik [15] employ clustering process to alleviate the data sparsity. Fundamentally, the reason for data scarcity problem is that inherent domain knowledge may be inaccurate and blurry. In general, the above work only considers the domain specific knowledge as prior information for data scarcity problem.

Generally, transfer learning has an advantage over the cold start problem in service recommendation [16]. It is designed to overcome the problem that machine learning has strict requirements on the size of data, the distribution of feature spaces and the integrity of labels especially for supervised learning [17]. It means that transfer learning could solve this problem with the help of other externally distributed data efficiently [18]. In addition, machine learning lacks generalization capabilities in new situations while the ability to transfer knowledge learned elsewhere to a new scene is the transfer learning. Among the CF based method, Zhang et al. [19] provides users with functionality based services rather than QoS based services [20]. In [21], *latent dirichlet allocation* (LDA) is employed to extract user-implicit-demand-factors and then generates a service recommendation list for users. In [22], their algorithm bases on the CF, firstly alleviates the sparsity problem with a novel ontology-based clustering approach, then determines trust value between users by calculating the correlation between users, finally obtains user rating prediction.

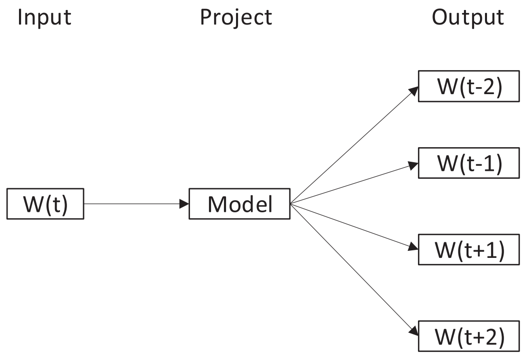

Fig. 2. The overview of skip-gram word2vec model.

In this paper, we present a system based on transfer learning to alleviate the cold start problem. The system is designed to overcome the problem of the rare and inaccurate descriptions of web services. It introduces additional knowledge from a general domain and specially reduce it to alleviate the data scarcity problem. In summary, the proposed work applies transfer learning to generate word2vec on service descriptions and fuses it with location information to obtain more accurate features from the source domain to the target domain. The introduced external domain experience is expected to solve cold start problem comparing with other mentioned methods from the filtered prior knowledge. As the location information will be fused with content as a vector to describe the similarity and try to give a rough prediction before MF, it considers that the rating matrix could be decomposed into two matrices with latent topics. From the experiments, it can be found that the transfer learning based word2vec model shows the potentiality to explore the relationship between topic words, and improve the LDA algorithm from the content relationship.

Overall, the main contributions of this work are: (1) The fusion method of content and location information for service recommendation with the help of transfer learning. (2) A new idea to initialize MF matrices and predict the missing values with *adaptive moment estimation* (ADAM). (3) The related comparative experiments with other state-of-art methods in both cold start and warm start environments. It can bring a balance between the sparse data to keep cloud security and the accuracy of the service recommendation.

2. Related work

2.1. Basic model of transfer learning

In general, transfer learning is widely employed in machine learning [23]. The basic assumption is that training process and future data are in the same distribution so that most tasks or data are related in transfer learning. Then if transfer learning works successfully, the performance would be improved by avoiding training models starting from scratch [16], which gives the mathematical definition of transfer learning and divide it into more detailed aspects by the dimensions of transfer learning settings, related areas, source domain labels, target domain labels and tasks. For the image data, taking the pre-trained deep learning models [24,25] to handle the large and challenging image classification is a common decision. While in the *natural language processing* (NLP) area, the models are suitable to be employed for content related tasks [26]. As shown in the Fig. 2, the skip-gram model predicts context with input word $w(t)$. After one-hot encoding for words, this model set a windows size sliding to generate input key-value pairs as input data. Then the word2vec completes training linear neurons as hidden layer and softmax classifier as output layer. Any word after hidden layer is projected into vector space as a word vector. The word vector, as an important feature, implies relationship in the context. As a powerful toolkit in NLP, it is possible to introduce transfer learning to process the raw data in service recommendation with word2vec.

In [23], the Deepmind team trains a robotic arm to move in the simulation environment, and then applies the progressive neural networks to the real robot arm. As a result, the real robotic arm can be trained to achieve the same effect as the simulation. The mainly ideal of progressive networks is to transfer the trained model weight to a new layer with a new task by the way named lateral connection. For the multi-task learning, Misra et al. [27] proposes a new sharing unit named “cross-stitch” combing the activations from multiple networks and can be trained end-to-end. The cross-stitch networks traverse all branches and points out the effect of different tasks on different forks. A similar idea also appears in [28]. The biggest difference between the cross-stitch network is that this work does not divide the different tasks into different network branches, but uses a single network structure to make the predictions sparse linear combination of residual units.

In service recommendation, comparing with image data, the text data contains high level features and is easier to be collected. It means the RS may need more careful analysis. Especially in cold start environment, the information about users or services is pretty sparse. So, it is essential to introduce transfer learning for extracting semantic information with word2vec in RS. From the above related work, the service recommendation performance is potential to be improved due to user characteristics could be modeled as multi-task learning.

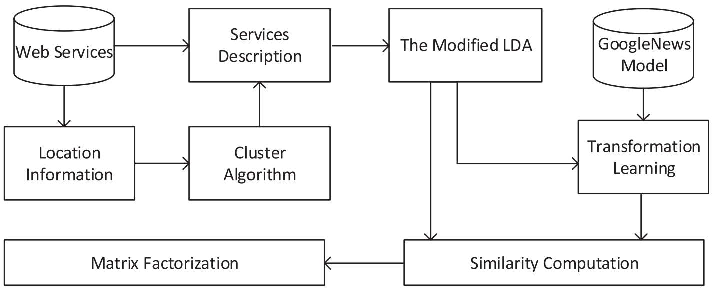

Fig. 3. The system overview.

2.2. Concepts of service recommendation

In the last decade years, the web services have taken the great changed on the world wild web [29]. With the exponential growth of web services, service recommendation ranges widely from kernel algorithms, data organization formats to service functions and so on. The least related work could be divided into two categories, one for the traditional methods represented with CF or its derivatives, the rest are based on the other machine learning algorithms.

The related methods could be further subdivided according to the directions of problems or the results of services. In [30], they propose a multi-label *active learning* (AL) with a correlation-aware learning strategy to address the challenges that tags annotated by users may be both inaccurate and incomplete. The AL starts from a very small training set and then selects the most informative services in the rest of services pool, asks a domain expert to tag them in iteration. The correlation-aware active approach learns the correlative information to further reduce experts' tagging effort. Compared with [21], they all ask for labels of the most informative services in iteration body, the difference is that the [21] takes SVM as base classifier and incorporates probabilistic topic models to improve performance. Wang et al. [31] defines five features based on the ranking result which produce the most effects on 960 real services. Also based on the content information, Liu et al. [32] starts with a base classifier and iteratively asks for the labels of the most informative services outside of the initial training set. Zhu and Cao [33] optimizes the metric *group-oriented mean average precision* (GMAP) based on the advantage of MAP to build the top-N applications list. It tries to model the latent relationship within cascaded characteristics in GMAP and recommends users' favorite applications.

The above approaches must fact the situation when users are newly to the RS and then recommend services based on the data scarcity problem. This data scarcity problem is defined in [11], called as the cold start problem.

It is a reasonable assumption that users in the near location tend to experience the similar QoS when they invoke the web services because local users may share the same network infrastructure. Lo et al. [11] proposals the local-based regularization on the classical popular MF to fill the missing values under the above assumption. Lee et al. [12] develops location-based MF via PP. The PP models a bipartite graph and applies constructs a random walk on this graph, and the preference could be propagated during this procedure. It tries to consider the location information as a regularization restraint in MF and force the model to perform recommendation based on the similar location groups. Bases on the CTR, Bai et al. [34] takes the generative process into account, finally models the procedure of service selection to deal with the cold start problem. Xie et al. [13] defines hidden correlation with asymmetric correlation matrices and applies random walking algorithm considering that the cold start problem is to process the sparse matrix between users and services in essence. The asymmetric correlation ranks of users and services are produced respectively. Finally, Top-K high ranked values are select to employ unified MF models. Yuan et al. [14] applies clustering process to users and service for reserving more information in next CF method. Similarly, Rupasingha and Paik [15] employs a novel ontology-based clustering to complete the non-rated data. Then it finds similar users and calculates trust weight values. Based on the result of trust weight values and user rating, it makes prediction for target user.

In this study, our proposed work is based on location information and MF. Similar with [35], this work also explores inherent contextual information to model RS. But [35] tends to mixture content-based and demographics-based model in RS. This work mainly takes advantage of location and content information to mine deep-seated features. In addition, comparing with the similar location-based method [12], our work regards location information as a special feature of services and users. Other methods like [36,37] try to simply obtain the decomposed matrices while this work aims to estimate a reasonable and credible decomposed matrix with the topic model at the initial stage.

3. Framework overview

3.1. Model overview

The overview system is shown in the Fig. 3.

In this model, the modified LDA components coarsely extract the topic association for each service. This unsupervised algorithm will extract high-level topics without any manual annotation. In real world, the large scale labeled data is hard to obtain, so the unsupervised algorithm is suitable for generic situation. Besides above reason, the LDA algorithm has been successfully applied to analysis various content data. But the non-content information: location also plays an important role in data source about service recommendation. It comes from the observation that similar area users will invoke similar services. The similar area refers that the adjacent relationship in geographical distance. In our work, the range of similar area depends on the parameter of cluster algorithm. In general, with the increase of cluster numbers, the range of similar area will gradually shrink and vice versa. This work attempts to obtain a scale factor between topic words and location information then try to fuse them in the modified LDA algorithm. Although LDA is a great algorithm for topic-modelling, it still has some limitations, mainly due to the fact that LDA does not consider position of the words in document so that extracted topic words may be overlapping in underlying semantics. In addition, in cold start environment, the size of data will be rare. So the naive LDA model may not work perfectly as people expect. Transfer learning and word2vec model is designed to alleviate above limitations. The primary word2vec model, which is trained with GoogleNews database in skip-gram architecture, performs word embedding operation for topic words after fine-tuning. It is a remarkable fact that the word2vec considers the positions of topic words in documents and further reveals the relationship between different topic words in vector dimension. In addition, this work considers introducing the negative sampling for the word2vec model. This technology is based on the word2vec mechanism that it finally uses the feature rather than the classification result in logistic regression classifier. In word2vec loss function, the negative sampling introduces the words expected value of probability of occurrence with log form from noise datasets. The meaning of loss function is to assign high probability to the real target words, and low probability to noise words. In related experiments, this method shows the ability to reduce the over-fitting problem in word2vec.

Based on the modified LDA, word2vec model projects topic words to high dimension space. Then, the combination of LDA and word2vec models will gradually mines semantic information to accurately describe users and services. The similarity computation component obtains correction weight and applies combination to compute the fine results and then generates initialization matrices for MF. MF predicts the missing value with weight decay operation and ADAM optimization algorithm. It utilizes that ADAM optimization algorithm will update gradient according to the sparse of parameters and works better in cold start problem.

In summary, this research problem is to model a service recommendation by using combing different components and taking advantage of them to gradually obtain precise result and alleviate data scarcity problem. The TLMF, our proposed algorithm, has more potentials to recommend service comparing with other state-of-art methods in cold start environment.

3.2. Model definitions

This section explains related definitions in web service, cold start problem and the TLMF. The Fig. 3 shows overview architecture and interaction relationship between several key components. In order to introduce detailed content, this work needs to give the definitions of related concepts.

The web service, as a kernel concept in RS, is a service offered by an electronic device to another electronic device, communicating with each other via the world wide web. The web service interface could be defined in *web services description language* (WSDL) format that includes essential information such as functions, parameters and return values. Based on the concept of the web service, the mashup is a kind of web service supporting different web *application program interfaces* (APIs).

In the modified LDA algorithm, the parameters *s, S, U, l* and *h* defines the input and output.

Definition 1. Due to the related data in mashup, this work describes the service as the tuple $s = \{id, name, tags, category, url\}$.

Definition 2. The term $S = \{s_1, s_2, s_3, \ldots, s_i\}$ refers to a set of services which are regarded as documents in LDA where $s_x (1 \le x \le i)$ denotes one service.

Definition 3. $U = \left\{u_1, u_2, u_3, \ldots, u_j\right\}$ refers to the set of users where $u_x (1 \le x \le j)$ denotes one user.

Definition 4. When the user u_i invokes the services $s_1, s_2, \ldots, s_k$, the location information l will be recorded and utilized in the following algorithm.

Definition 5. The set $h_i = \{s_1, s_2, \ldots, s_k\}$ represents the user i history data and the cardinality of this set is k.

In mathematics, this model takes the service description d filtered from s and location based information l as input data for the modified LDA model. It will generate the topic association $\vec{z}$. As the most important output of the modified LDA, $\vec{z}$ will be jointly used with S in transfer learning model.

Definition 6. The output of the modified LDA includes the topic association $\vec{z}$, topic-word probability distribution Θ, topic-document probability distribution ϕ.

The transfer learning model takes the topic association $\vec{z}$ and service description d as input data. The output of transfer learning model is the correction coefficient w.

Definition 7. The w is weight correction coefficient for topic word $\vec{v}$. As the output of transfer learning, w is used to perform data combination to obtain the initialization value for MF.

The MF takes the corrected result after initialization as scoring matrix $R_{m,n}$. The goal of naive MF is to predict the decomposed matrix P and Q in service.

Definition 8. The $R_{m,n}$ is a highly sparse matrix where the m and n means users and topic separately. The matrix $R_{m,n}$ is the product of matrix $P_{m,k'}$ and matrix $Q_{k',n}$. The $r_{u,i}$ represents the degree of the user u interest in the service s. The matrix $\hat{R_{m,n}}$ is the reconstructed matrix that the missed value will be predicted by P and Q.

$$R \approx P \cdot Q = \hat{R} \tag{1}$$

In cold start problem, the cardinality of set h will be small. This situation means that the naive MF based method will have to handle sparse matrix without any pretreatment.

Definition 9. According to a definition provided by [38], S represents the sparseness measure based on the relationship between the L_1 norm and L_2 norm as:

$$\frac{\sqrt{n} - (\sum |x_i| / \sqrt{\sum x_i^2})}{\sqrt{n} - 1} \tag{2}$$

where the n is the dimensionality of x.

It will be used to evaluate the sparsity of R in experiment section.

3.3. Program descriptions

In the modified LDA, there are some essential definitions to description this algorithm. In order to introduce the location information into the modified LDA algorithm, the proportion relationship after introducing between top-n location count l and top-n word count f_{wd} is defined as:

Definition 10. Assuming that the frequency of word is f_{wd} in service and the f_l is the frequency of location, then the f_{wd} and f_l are in proportion relationship as:

$$\frac{\sum f_l}{\sum f_{wd}} = \gamma \tag{3}$$

where the $\gamma = \max\{\min\{\omega + \sum l_i, \eta\}, \delta\}$. The parameter ω, η and δ are given in program.

The transfer learning model is employed to transfer known knowledge.

Definition 11. Given that the source domain D_s and learning task T_s, transfer model f projects the D_s to target domain D_r: $f(D_s)$.

Definition 12. The negative words could be selected with the experience empirical formula:

$$P(w_i) = \frac{f(w_i)^{\xi}}{\sum_{j=0}^{n} f(w_j)^{\xi}} \tag{4}$$

where the $f(w)$ represents the frequency for the word w and the ξ is $\frac{3}{4}$ as recommendation.

Definition 13. The negative sampling defines the loss function as:

$$E = -\log \sigma(v'_{w_O} h) - \sum_{w \in W_{neg}} \log \sigma(-v'_w h) \tag{5}$$

where the v'_{w_O} is the output vector, the W_{neg} is the negative sampling dataset, σ is the sigma function.

Definition 14. For any vector $\vec{v_j}$ and $\vec{v_k}$ in set $\vec{v}$, the similarity between them could be measured as:

$$dis(\vec{v_j}, \vec{v_k}) = euclidean(\vec{v}_j, \vec{v}_k) = \sqrt{(\vec{v_j} - \vec{v_k})^T (\vec{v_j} - \vec{v_k})} \tag{6}$$

Definition 15. The weight w for each topic word $\vec{v}$ could be calculated:

$$w = 1 - \frac{m \cdot dis(\vec{v}, \vec{center}) - \sum_{\vec{v'} \in cluster} dis(\vec{v'}, \vec{center})}{m \cdot R} \tag{7}$$

where m is the number for the words in $\vec{v}$ related cluster, the parameter R is the maximum distance in these cluster pairs:

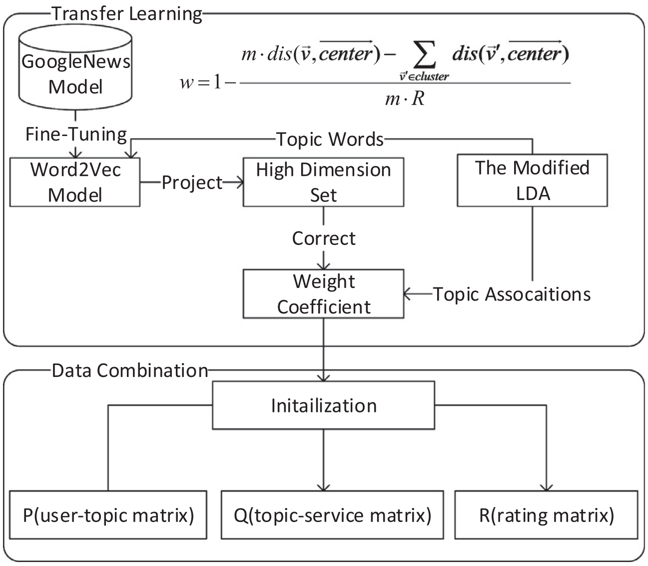

Fig. 4. The detail progress shows the data flow from transfer learning to MF. The transfer learning projects the modified LDA topic to high dimension space and then performs correction on origin data. Then the combination results will be used as the initialization values in MF.

$$R = \max_{\vec{v}' \in cluster} dis(\vec{center}, \vec{v}'). \tag{8}$$

Definition 16. The predicted value could be written as $\hat{r_{u,i}} = \sum_{f=1}^{F} P_{u,f} Q_{f,i}$.

Definition 17. The MF optimized the loss function:

$$min: J(r) = \sum ((r_{u,i} - \hat{r_{u,i}}) \cdot dw_{t,u,i})^2 + \lambda (\sum P_{u,f}^2 + \sum Q_{f,i}^2) \tag{9}$$

where the λ is regularization parameter given manually, and the parameter $dw_{t,\, u,\, i}$ is the weight decay coefficient which will be update with step t and decay rate dr:

$$dw_{t,u,i} = \begin{cases} 1, & r_{u,i} \text{ is not missing value} \\ exp(-dr \cdot t), & r_{u,i} \text{ is missing value} \end{cases} \tag{10}$$

Definition 18. The user-topic matrix P will be initialized as:

$$P_i = \sum_{s_j \in h_i} r_{i,j} \cdot f(s_j) \tag{11}$$

where the $f(s_j)$ is the accumulative services scores determined by therein topic words:

$$f(s_j) = \sum_{d \in s_j} \Phi_d \sum_{\vec{z} \in d} \vec{z} \cdot w \tag{12}$$

Definition 19. The topic-item matrix Q will be initialized as:

$$Q = \phi^T \tag{13}$$

Definition 20. The missing value in rating matrix R will be determined as:

$$R_{i,j} = \begin{cases} \Phi_j \cdot avg(\sum_{s_k \in h_i} \frac{r_{i,k} \cdot f(s_k)}{N_i}), & N_i = 0 \\ \sum_{u_k \in U} R_{k,j}, & N_i \neq 0 \end{cases} \tag{14}$$

where the N_i refers the number of services in set h_i.

4. Detailed algorithm

4.1. The modified LDA

This component mainly analyses content-based and location information in the modified LDA. It is an unsupervised clustering algorithm to compute the most related topic for each existing service. It projects the service description d to a vector of terms composed of multiple topics. In addition, this work adopts the cluster algorithm for location information and obtains its frequency. The location information l is regarded as a specific word in service description, while each service description only preserves top-n frequency as the prior probability. After processing one-hot encoding for l, this work averagely scatters l in d with the Eq. 3.

In experience, it is hard to directly estimate parameters for the whole dataset with original LDA maximum formula. The Gibbs Sampling is employed to complete approximate evaluation. A simple process described in pseudo code could be used to complete the primary processing shown in Algorithm 1. Besides above parameter, the input of the modified LDA includes the hyperparameter α, β and topic number K. By the end of sampling convergence, multinomial topic-word probability distribution matrix Θ, multinomial document-topic probability distribution matrix ϕ and topic association $\vec{z}$ are obtained.

Algorithm 1 The modified LDA model with Gibbs sampling.

Input: service description d, hyperparameter α, β, topic number K, location information l
Output: topic association $\vec{z}$, multinomial parameters ϕ and Θ

averagely apply k-means to l and apply filtering
zero all count variables, n_m^k, n_m, n_k
for all documents $m \in [1, M]$ **do**
    scatters l_i in m.
    for all all words $n \in [1, N_m]$ in documents m **do**
        sample topic index $z_{m,n} = k \sim Mult(1/K)$
        increment document topic count: $n_m^k = n_m^k + 1$
        increment document topic sum: $n_m = n_m + 1$
        increment topic term count: $n_k^t = n_k^t + 1$
    end for
end for
while not finished **do**
    for all all documents $m \in [1, M]$ **do**
        decrement counts and sums: $n_m^k = n_m^k - 1$; $n_m = n_m - 1$, $n_k^t = n_k^t - 1$; $n_k = n_k - 1$
        sample topic index
        increment counts and sums: $n_m^k = n_m^k + 1$, $n_m = n_m + 1$, $n_k = n_k + 1$
    end for
    if converged and L sampling iterations since last read out **then**
        read out parameter ϕ
        read out parameter Θ
    end if
end while

4.2. Word2vec model based on transfer learning

This component mainly applies transfer learning model to correct the coarse result from the modified LDA. After the modified LDA algorithm, the topic words are projected into high dimension space with transfer learning. This work assigns the D_s with the GoogleNews model and the D_r with service description content. It means that the source target D_r is trained with the data collected in GoogleNews. The GoogleNews is a pre-trained word2vec model with Google News corpus. This work assumes that there exists similarity between the service description content domain D_t and the source domain Google News D_s. This progress is illustrated in the Fig. 4. The modified LDA provides the specific word $\vec{z}$ for model f. Then the topic word after transfer model is $\vec{v}$. It plays an important role in correction, data combination to improve the performance of the modified LDA by introducing GoogleNews model knowledge.

In general, fine-tuning is employed to retrain the learning model under the GoogleNews model. In order to boost the convergence speed and avoid over-fitting in fine-tuning, this work applies the negative sampling. It will optimize the loss function E with Eq. (5). Eq. (4) will select the high frequency words as the negative words. In the modified LDA, the added location information will have an effect on the frequency of location word. After the negative sampling, this work generates the negative sampling set with negative words. Assuming that the target user invokes n services, after the LDA algorithm, the related services belong to k topics. For the target user, the n services s involve k' topics t that could be written as the form: $\{t \mid t \ in \ s_i, i = 1, 2, \ldots, n\}$, where the number of the set is k'. So the content information t employs transfer learning with

$\vec{v} : f(t)$, where the $\vec{v}$ is a high dimension vector set. For any vector $\vec{v_j}$ and $\vec{v_k}$ in set $\vec{v}$, the similarity between them could be measured with Eq. (6). Based on the following similarity measure formula, this work applies k-means cluster algorithm then obtains the clustered center point $\vec{center_i}$ for each topic vector in $\vec{v}$. At the same time, the weight w for each topic word could be calculated as Eq. (7). A simple process described in pseudo code can be used to complete the processing for transfer learning is shown in Algorithm 2.

Algorithm 2 Word2vec model based on the transfer learning.

Input: model m trained based on the GoogleNews, topic association $\vec{z}$, service description d, topic word $\vec{v}$, k-means parameter K and involved topic set t
Output: weight correction coefficient w_i for topic word $\vec{v_i}$

```
for all all words w in service description do
    compute the frequency P(w_i) of word w_i
end for
apply the negative sampling to select the negative words set W_neg with Equation 4.
apply the fine-tuning to the model m with service description d and negative words set W_neg then obtain the transfer learning model f
for all topic word of description v_d of service description d do
    feed the topic word |z| to f and obtain the projected vector v
end for
apply k-means cluster with parameter K under the similarity measure Equation 6 and obtain the center vector center
for all topic word in set t do
    compute the weight correction coefficient w with Equation 7.
end for
```

4.3. Matrix factorization

This component accepts the target user request and collects the location based information, then makes prediction via MF. In theory, the matrix P will be regarded as user-topic matrix, while the matrix Q is the topic-service matrix. In naive CF based methods, P and Q will be initialized randomly. But this work determines the P, Q and R with the Eq. (11), Eq. (13) and Eq. (14). These steps aim to give a better initialization for parameter P and Q and try to make approximate prediction for missed value based on the result of LDA model and transfer learning. The related experiments show that the better initialization will speed up the progress of convergence. The weight decay operation aims to preserve the predicted values by Eq. (14) because it provides interpretable results at the initial stage. But the predicted result for missing values may not be so exact comparing with other origin ratings records, so this MF needs to decay the influence of them to avoid over-fitting.

This is a highly sparse matrix due to the cold start problem. To alleviate data scarcity problem, this work add the regularization term for the loss function Eq. (9). In addition, the ADAM optimization algorithm works better for sparse data in related experiments. It could adapt the learning rate to the parameters, performing larger updates for infrequent and smaller updates for frequent parameters. Especially in cold start problem, the input data for matrix factorization is sparse. A simple process described in pseudo code can be used to complete the processing for MF is shown in Algorithm 3.

Algorithm 3 Matrix factorization.

Input: weight correction coefficient w, topic association $\vec{z}$, document-topic probability distribution Φ, regularization parameter λ, parameter f and weight decay rate dw
Output: the decomposed $P_{u,f}$, matrix $Q_{f,i}$ and matrix $\hat{R_{u,i}}$

```
initialize the user-topic matrix P_{u,f} with Equation 11.
initialize the topic-service matrix Q_{f,i} with Equation 13.
initialize the miss value in rating matrix R_{u,i} with Equation 14.
for the optimization objective function Equation 9, apply the ADAM optimization algorithm
obtain the decomposed P_{u,f} and matrix Q_{f,i}
obtain the reconstructed matrix R^_{u,i} with Equation 10 and recommend top-n services with euclidean distance
```

5. Experiments and results

This section presents experiments to evaluate performance. First, we need to introduce the Book-Crossing [39] dataset. Then, we discuss the results of discovering individual service topics in the system. At last, we show that TLMF over-performs baselines when making service recommendation.

Table 1
Topics model.

topic 1	topic 2	topic 3	topic 4	topic 5
war	work	love	canada	united
world	classic	story	free	kingdom
american	history	man	victim	page
soldier	world	husband	murder	shows
fiction	include	woman	blood	cover
usa	edition	family	red	printing

Performance evaluations of the prediction results involve *Mean Absolute Error* (MAE) and *Root Mean Square Error* (RMSE) in a comparison with previous approaches. They generate impact of accuracy for RS in Book-Crossing dataset. We compare the error rate of different recommendation results with different clustering approaches for the sparsity alleviating.

5.1. Dataset description

The service recommendation performance needs to be evaluated with the real word large dataset. This dataset is collected by Ziegler et al. [39] from the Book-Crossing community. It contains 278,858 users (anonymized but with demographic information) providing 1,149,780 ratings (explicit / implicit) about 271,379 books.

Besides, in order to analysis topic, this work crawls brief introduction of involved books in Amazon. It means that this experiment mainly utilizes the users rating information, brief introduction and users location information. After simply filtering for origin data set, we obtain the filtered user-service rating matrix with 11,309 rows, 7124 cols and 34,799 rating records. The sparsity of this user-service rating matrix is up to 0.9879. This work projects the origin ratings data to [0.000001, 1].

As discussed above, the user-rating matrix is really sparse, this work explores the performance between different sparsity with different recommendation algorithms.

5.2. Model topics analysis

In this experiment, the topics of documents represent the most important features. In order to extract topic model in the modified LDA algorithm, we mine content information, focusing on taxonomic descriptions of books in Amazon.com. This experiment regards the brief introduction to books as the service description.

This work performs the modified LDA with 10 latent topics, $\alpha = 2.0,\ \beta = 0.5,\ \omega = 0.005,\ \eta = 0.01$ and $\delta = 0.25$.

There are some selected topics with top 6 terms shown in Table 1. These terms have the highest probability of belonging to the corresponding topics. As shown in Table 1, the modified LDA has coarsely extracted meaningful topics words belonging to different fields.

From the Table 1, the modified LDA algorithm has fused the location information with origin text content in topic 1, topic 4 and topic 5.

5.3. Transfer learning analysis

Transfer learning based word2vec could efficiently reveal the relationship between different words. In our detailed experiment, the dimensions of vectors are all 300 to represent words. The original GoogleNews model's words vectors are regarded as the model initialization values in transfer learning. After t-SNE processing for the result of word2vec, this work randomly samples 2D points and applies cluster algorithm. As shown in Fig. 5, the visualization result is adopted with K-means algorithm with 10 clusters to analyze similarity between the nearest words vectors.

The 2D visualization result shows that words with similar semantics still keep the adjacent relationship of this space. On the one hand, it is the mapping result from the origin space. It means that in high dimension space, the related words vectors still hold the similar relationship as in visualization result. In visualization result, the fused location information 'usa' keeps the adjacent relationship to the topic words 'american'. Besides that, the relevant semantic words like 'woman', 'mother', 'love' and 'husband' hold close distances in the cluster colored with yellow.

On the other hand, it shows the potential for transfer learning in dealing with data scarcity, especially under the situation that valuable information about collected contents of books is rarely. The average length for each book description are just about 32.77 before the simply processing steps including stemming, lemmatization, deleting stop words etc. The origin topic model may not exactly describe the complete book because brief description usually represents two or three major features to attract readers. With the help of transfer learning, after crawling about 2.63% brief introductions of books, this work successfully outperforms other state-of-the-art techniques.

5.4. Comparative experiment

After fine-tuning process, this work applies ADAM optimization to rating matrix with default hyperparameters.

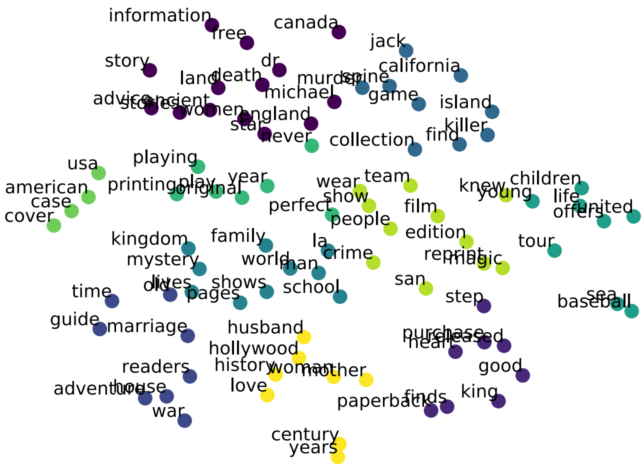

Fig. 5. The visualization result after t-SNE algorithm.

Table 2
Cold start environment.

METHOD	MAE			RMSE		
	10%	20%	30%	10%	20%	30%
SVD	0.3359	0.3260	0.3189	0.3745	0.3711	0.3691
SVD++	0.3292	0.3190	0.3116	0.3704	0.3661	0.3642
NMF	0.3259	0.3114	0.3030	0.3895	0.3897	0.3913
k-NN	0.3561	0.3521	0.3471	0.3871	0.3946	0.4009
Slop One	0.3290	0.3205	0.3131	0.3913	0.3944	0.3921
Baseline	0.3421	0.3336	0.3269	0.3734	0.3682	0.3658
TLMF	0.2722	0.2721	0.2704	0.3355	0.3347	0.3424

Table 3
Warm start environment.

METHOD	MAE			RMSE		
	60%	70%	80%	60%	70%	80%
SVD	0.3096	0.3049	0.3026	0.3657	0.3604	0.3586
SVD++	0.3297	0.3192	0.3116	0.3710	0.3656	0.3644
NMF	0.2888	0.2842	0.2815	0.3903	0.3871	0.3790
k-NN	0.3364	0.3334	0.3265	0.4060	0.4036	0.3973
Slop One	0.3017	0.2968	0.2937	0.3892	0.3854	0.3798
Baseline	0.3164	0.3135	0.3099	0.3600	0.3572	0.3538
TLMF	0.2682	0.2674	0.2668	0.3477	0.3347	0.3309

In this section, the MSE and RMSE is used as the evaluation metric to study the performance of different methods. MSE is a measure of the average absolute deviation between a predicted rating and the user's true rating. It forms as $MAE = \frac{\sum_j^N |r_i - \hat{r}_i|}{N}$, where the N is the real number of real value r_j. The RMSE measures the errors of the predicted rating by quadratic scoring the differences between prediction values and true ratings: $RMSE = \sqrt{\frac{\sum_j^N |r_i - \hat{r}_i|^2}{N}}$.

To evaluate this framework over different sparsity, this work samples different sizes of training data to represent the cold start environment. They respectively cover 10%, 20% and 30% of total dataset. In addition, the sizes of training data that are 60%, 70% and 80% of a total dataset indicates the warm start environment. The training data are randomly selected from the total data, and the rest data in the matrix are used for test data.

After scaling origin data, seven service recommendation algorithms are compared in this experiment and results are shown in Tables 2 and 3. In general, the memory-based CF methods usually work better than the numerical approaches, and model-based CF presents the higher prediction accuracy than the memory-based CF in related experiments [12]. So this work mainly compares our work with model-based CF methods.

To evaluate the proposed method, it should be compared with the existing state-of-the-art techniques. Implemented techniques for the comparison are SVD [37], SVD++ [40], NMF [36], Slop One [41], k-NN and Baseline [42]. We have found several preliminary results from the Tables 2 and 3. In general, the prediction of all comparison techniques are improved

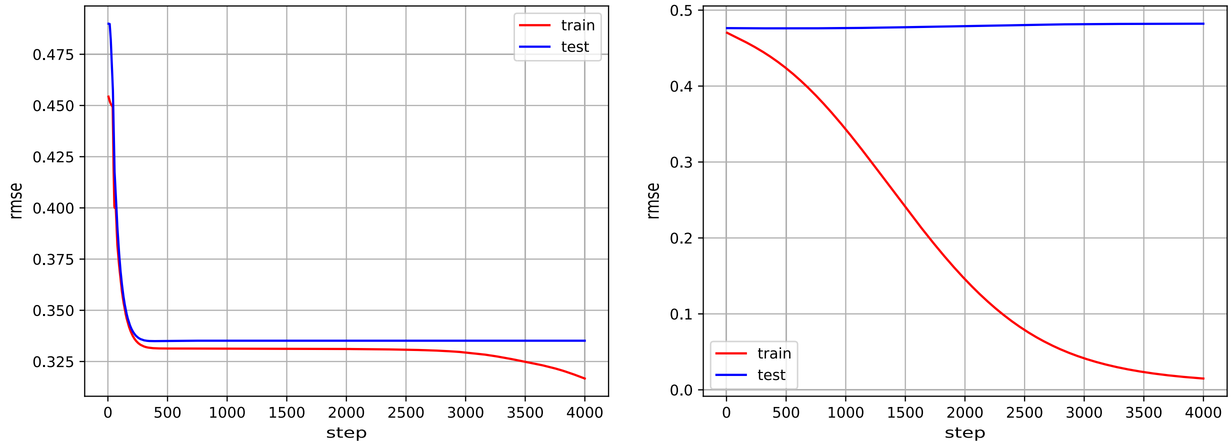

Fig. 6. These figures show the performances between our proposed TLMF method(shown in left) and naive MF method(shown in right) without matrices initialization and weight decay.

with the increase of the matrix density. In these techniques, our work, the TLMF shows the best performance with the lowest MAE and RMSE. It means our work could give a better result in both cold start environment and warm start environment. Besides that, our method shows the most stable performance.

In addition, the matrices initialization and weight decay operations are very important technologies to speed up the convergence of loss function and alleviate the over-fitting problems. In specific experiments with the same parameters, the naive MF usually converges more slowly than our proposed TLMF. When we remove the weight decay operation, the naive MF may cause over-fitting in test dataset. As shown in the Fig. 6, our proposed TLMF finally converges within about 300 steps while the naive MF still processes training and gets stuck into over-fitting till training done. From this aspect, the transfer learning is proved to be effective when facing the cold start problem.

6. Conclusions and future work

This work presents transfer learning based framework which employs other domain knowledge, fusing with content and location information to provide robust recommendation results. The main goal of this work is to determine the performance of transfer learning in data scarcity problem. The second aim of this work is to investigate the effects of combination of LDA and transfer learning. The transfer learning provides an environment for the modified LDA to utilize similar domain knowledge to improve framework performance. Comparing with other model-based methods, the main idea of this approach is introducing location information and another domain knowledge.

From the result of experiments, the word2vec model shows the potentiality to explore the relationship between topic words. It is revelatory to improve the naive LDA algorithm from content relationship. This study has found that the transfer learning could efficiently alleviate data scarcity with matrices initialization and weight decay. With the little source domain knowledge, transfer learning successfully mines latent semantic information to describe the relationship between topic words. In both cold start environment and warm start environment, the proposed algorithm TLMF is more robust than other model-based state-of-art methods.

In the future, our work will focus on processing large dataset and improving the system performance. It strictly requires that the input data with content and location information. So, we will explore the method to reduce restrictions of data and improve performance in large dataset. Besides, this research has thrown up many questions in need of further investigation. Considerably more work will need to be done to determine concrete result on large dataset. The issue of the choice of source domain is an intriguing one which could be usefully explored in further research. We will explore the relationship between source domain and service recommendation. More information on similarity between source domain and target domain would help us to establish a greater degree of accuracy on this matter.

Declaration of Competing Interest

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.

Acknowledgments

This work has been partially supported by the National Key Research and Development Program (No. 2018YFB1402800), the project of the Independent Innovation Engineering of Shandong Province (No. 2017CXGC0703).

References

[1] L. Wu, S.K. Garg, S. Versteeg, R. Buyya, Sla-based resource provisioning for hosted software-as-a-service applications in cloud computing environments, IEEE Trans. Serv. Comput. 7 (3) (2014) 465–485.
[2] Y. Fanjiang, Y. Syu, S. Ma, J. Kuo, An overview and classification of service description approaches in automated service composition research, IEEE Trans. Serv. Comput. 10 (2) (2017) 176–189.
[3] Y. Hao, Y. Fan, W. Tan, J. Zhang, Service recommendation based on targeted reconstruction of service descriptions, in: 2017 IEEE International Conference on Web Services (ICWS), 2017, pp. 285–292.
[4] K. Kritikos, D. Plexousakis, Requirements for qos-based web service description and discovery, IEEE Trans. Serv. Comput. 2 (4) (2009) 320–337.
[5] W. Lo, J. Yin, Z. Wu, Accelerated sparse learning on tag annotation for web service discovery, in: 2015 IEEE International Conference on Web Services, 2015, pp. 265–272.
[6] G. Sun, T. Cui, D. Xu, H. Chen, S. Chen, J. Shen, Assisting open education resource providers and instructors to deal with cold start problem in adaptive micro learning: aservice oriented solution, in: 2017 IEEE International Conference on Services Computing (SCC), 2017, pp. 196–203.
[7] J. Li, J. Wang, Q. Sun, A. Zhou, Temporal influences-aware collaborative filtering for qos-based service recommendation, in: 2017 IEEE International Conference on Services Computing (SCC), 2017, pp. 471–474.
[8] P. Zhang, Y. Liu, M. Qiu, Snc: a cloud service platform for symbolic-numeric computation using just-in-time compilation, IEEE Trans. Cloud Comput. 7 (2) (2019) 580–592.
[9] K. Gai, M. Qiu, H. Zhao, Security-aware efficient mass distributed storage approach for cloud systems in big data, in: 2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS), 2016, pp. 140–145.
[10] M. Qiu, W. Dai, A.V. Vasilakos, Loop parallelism maximization for multimedia data processing in mobile vehicular clouds, IEEE Trans. Cloud Comput. 7 (1) (2019) 250–258.
[11] W. Lo, J. Yin, S. Deng, Y. Li, Z. Wu, Collaborative web service qos prediction with location-based regularization, in: 2012 IEEE 19th International Conference on Web Services, IEEE, 2012, pp. 464–471.
[12] K. Lee, J. Park, J. Baik, Location-based web service qos prediction via preference propagation for improving cold start problem, in: 2015 IEEE International Conference on Web Services, IEEE, 2015, pp. 177–184.
[13] Q. Xie, S. Zhao, Z. Zheng, J. Zhu, M.R. Lyu, Asymmetric correlation regularized matrix factorization for web service recommendation, in: 2016 IEEE International Conference on Web Services (ICWS), IEEE, 2016, pp. 204–211.
[14] Y. Yuan, W. Zhang, X. Zhang, Location-based two-phase clustering for web service qos prediction, in: 2016 13th Web Information Systems and Applications Conference (WISA), IEEE, 2016, pp. 7–11.
[15] R.A. Rupasingha, I. Paik, Improving service recommendation by alleviating the sparsity with a novel ontology-based clustering, in: 2018 IEEE International Conference on Web Services (ICWS), IEEE, 2018, pp. 351–354.
[16] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.
[17] M. Long, H. Zhu, J. Wang, M.I. Jordan, Unsupervised domain adaptation with residual transfer networks, in: Advances in Neural Information Processing Systems, 2016, pp. 136–144.
[18] M. Qiu, K. Gai, B. Thuraisingham, L. Tao, H. Zhao, Proactive user-centric secure data scheme using attribute-based semantic access controls for mobile clouds in financial industry, Future Gener. Comput. Syst. 80 (2018) 421–429.
[19] Y. Zhang, T. Lei, Y. Wang, A service recommendation algorithm based on modeling of implicit demands, in: 2016 IEEE International Conference on Web Services (ICWS), IEEE, 2016, pp. 17–24.
[20] Z. Zheng, H. Ma, M.R. Lyu, I. King, Qos-aware web service recommendation by collaborative filtering, IEEE Trans. Serv. Comput. 4 (2) (2011) 140–152.
[21] X. Liu, S. Agarwal, C. Ding, Q. Yu, An lda-svm active learning framework for web service classification, in: 2016 IEEE International Conference on Web Services (ICWS), IEEE, 2016, pp. 49–56.
[22] R.A. Rupasingha, I. Paik, Improving service recommendation by alleviating the sparsity with a novel ontology-based clustering, in: 2018 IEEE International Conference on Web Services (ICWS), IEEE, 2018, pp. 351–354.
[23] C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L.-J. Li, L. Fei-Fei, A. Yuille, J. Huang, K. Murphy, Progressive neural architecture search, 2018, pp. 19–34.
[24] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[25] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[26] J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.
[27] I. Misra, A. Shrivastava, A. Gupta, M. Hebert, Cross-stitch networks for multi-task learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3994–4003.
[28] C. Doersch, A. Zisserman, Multi-task self-supervised visual learning, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2051–2060.
[29] E. Al-Masri, Q.H. Mahmoud, Investigating web services on the world wide web, in: Proceedings of the 17th International Conference on World Wide Web, ACM, 2008, pp. 795–804.
[30] W. Shi, X. Liu, Q. Yu, Correlation-aware multi-label active learning for web service tag recommendation, in: 2017 IEEE International Conference on Web Services (ICWS), IEEE, 2017, pp. 229–236.
[31] S. Wang, Y. Zou, J. Ng, T. Ng, Learning to reuse user inputs in service composition, in: 2015 IEEE International Conference on Web Services, IEEE, 2015, pp. 695–702.
[32] L. Liu, F. Lecue, N. Mehandjiev, L. Xu, Using context similarity for service recommendation, in: 2010 IEEE Fourth International Conference on Semantic Computing, IEEE, 2010, pp. 277–284.
[33] N. Zhu, J. Cao, Gtrm: A top-n recommendation model for smartphone applications, in: 2017 IEEE International Conference on Web Services (ICWS), IEEE, 2017, pp. 309–316.
[34] B. Bai, Y. Fan, K. Huang, W. Tan, B. Xia, S. Chen, Service recommendation for mashup creation based on time-aware collaborative domain regression, in: 2015 IEEE International Conference on Web Services, IEEE, 2015, pp. 209–216.
[35] V.N. Zhao, M. Moh, T. Moh, Contextual-aware hybrid recommender system for mixed cold-start problems in privacy protection, in: 2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS), 2016, pp. 400–405.
[36] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, in: Advances in Neural Information Processing Systems, 2001, pp. 556–562.
[37] Y. Koren, R. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, Computer (8) (2009) 30–37.
[38] P.O. Hoyer, Non-negative matrix factorization with sparseness constraints, J. Mach. Learn.Res 5 (1) (2004) 1457–1469.
[39] C.-N. Ziegler, S.M. McNee, J.A. Konstan, G. Lausen, Improving recommendation lists through topic diversification, in: Proceedings of the 14th International Conference on World Wide Web, ACM, 2005, pp. 22–32.
[40] Y. Koren, Factorization meets the neighborhood: a multifaceted collaborative filtering model, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2008, pp. 426–434.
[41] D. Lemire, A. Maclachlan, Slope one predictors for online rating-based collaborative filtering, in: Proceedings of the 2005 SIAM International Conference on Data Mining, SIAM, 2005, pp. 471–475.

[42] Y. Koren, Factor in the neighbors: scalable and accurate collaborative filtering, ACM Trans. Knowl. Discov. Data (TKDD) 4 (1) (2010) 1–24.

Chao Lei received his Bachelor degree of Engineering at College of Software, Shandong University, China in 2018. His research interests include natural language processing and service recommendation.

Hongjun Dai received his BE and PhD in Computer Science from Zhejiang University, China in 2002 and 2007, respectively. Currently, he is an Associate Professor of Software Engineering at Shandong University, China. His research interests include reliability of the novel computer architecture such as multicore processor, cloud, mobile internet, etc., optimization of wireless sensor networks, modeling of the cyber-physical systems. He has published more than 30 peer-reviewed journal and conference papers, and earned seven Chinese patents. His research is supported by the National Science Foundation of China, the Chinese Department of Technology, and the companies such as Intel and Inspur.

Zhilou Yu received the B.E. degrees of information and electronic engineering from Zhejiang University, China in 1993, and the Ph.D. degrees of information and communication engineering from Shandong University, China in 2001, respectively. Currently, he is working at the Chaoyue Digital Electronics Co.Ltd, China. His research interests include embedded systems, wireless sensor networks, cyber-physical systems. He has published more than 50 peer-reviewed journal and conference papers, and earned 70 Chinese patents. His research is supported by the National Science Foundation of China, the Chinese Department of Technology.

Rui Li received the Ph.D degree of computer science from Technical University of Munich, Germany, in 2013. Currently, he is the chief AI scientist of AI Research Institute, Inspur Inc., China. Prior to joining Inspur Inc., he has worked for Siemens (Germany) and Alibaba (China) respectively. His research interests include data mining, machine learning and related applications in manufacturing and medicine. He has published more than ten peer-reviewed journal and conference papers as well as a book chapter published by the Springer. He is also a committee member of Artificial Intelligence and Pattern Recognition (CCF-AI) of China Computer Federation.