Wireless Communications and Mobile Computing

Natural Language Processing Empowered Mobile Computing

Lead Guest Editor: Tianyong Hao
Guest Editors: Raymond Wong, Zhe He, Haoran Xie, Tak-Lam Wong, and Fu Lee Wang

Natural Language Processing Empowered Mobile Computing

Wireless Communications and Mobile Computing

Natural Language Processing Empowered Mobile Computing

Lead Guest Editor: Tianyong Hao
Guest Editors: Raymond Wong, Zhe He, Haoran Xie, Tak-Lam Wong, and Fu Lee Wang

Editorial Board

Contents

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 9130545, 2 pages
https://doi.org/10.1155/2018/9130545

Editorial

Natural Language Processing Empowered Mobile Computing

Tianyong Hao,[1] **Raymond Wong**,[2] **Zhe He**,[3] **Haoran Xie**,[4] **Tak-Lam Wong**,[5] **and Fu Lee Wang**[6]

[1]*South China Normal University, Guangzhou, China*
[2]*University of New South Wales, Sydney, Australia*
[3]*Florida State University, Tallahassee, USA*
[4]*The Education University of Hong Kong, New Territories, Hong Kong*
[5]*Douglas College, British Columbia, Canada*
[6]*The Open University of Hong Kong, Hong Kong*

Correspondence should be addressed to Tianyong Hao; haoty@gdufs.edu.cn

Received 13 September 2018; Accepted 13 September 2018; Published 14 October 2018

With the rapid growth of mobile device usage, more and more mobile contents offer a great opportunity for mining useful information. However, these contents mostly exist in free-text format in web pages, news feeds, SMS, and Twitter/WeChat messages, posing a significant challenge for information extraction. Natural language processing (NLP) is an important field of computer science, artificial intelligence, and computational linguistics. It is concerned with the interactions between computers and human (natural) language. NLP aims to enable a computer program to process and understand unstructured texts. In the mobile environment, NLP can make apps smarter by automatically analyzing the content, understanding its semantics, and taking appropriate actions on behalf of their users. The state-of-the-art NLP techniques have proven to be useful in dealing with the information overload problem in the mobile environment, e.g., news aggregation and summarization, question answering, information extraction and retrieval, semantic understanding, and personalization.

Applying NLP techniques to the content on mobile devices has recently gained attention in both academia and industry. Efforts have been made to develop various NLP applications for mobile platforms, such as Siri, which can retrieve information about places or events, and Google Now, which can extract context-sensitive information. However, it still lacks deep text processing capabilities, such as abstraction, summarization, and semantic understanding to accomplish complex and information-intensive tasks. Beyond the common applications, in the fast growing areas, such as healthcare, critical electronic health record data are accessible almost anywhere and near instantly even in ICU. Patient notes can be retrieved with a few taps on a screen. Sophisticated NLP techniques are required for analyzing the huge volume of data timely for making critical decisions.

In this special issue, articles regarding the use of technologies, methodologies, and applications for natural language processing and mobile computing were submitted and peer-reviewed. The special issue shows a diversity of new developments in these areas. We accepted 7 high-quality original research articles and 1 review article.

The paper "A Bibliometric Review of Natural Language Processing Empowered Mobile Computing" introduced a systematical investigation and analysis on the scholarly publications on the topic of natural language processing empowered mobile computing in the last ten years. The authors applied a number of analytical techniques including descriptive statistics, geographic visualization, social network analysis, Latent Dirichlet Allocation, and affinity propagation clustering, to discover the status of research efforts and the trend of the topic. The paper can potentially help researchers understand research hot spots, collaboration patterns, and scholarly resource distribution as well as trace new scientific development in the field.

The paper "The Current Status and a New Approach for Chinese Doctors to Obtain Medical Knowledge Using Social Media: A Study of WeChat" provided a quantitatively exploration of the approaches to acquiring medical knowledge using social media such as WeChat. Focusing on WeChat, the most widely used mobile social media in China, the authors designed, distributed, collected, and analyzed a self-administered questionnaire utilizing an online survey tool. The paper reported the most desirable mode for acquiring professional medical knowledge through WeChat from data analysis. The paper advocates both academia and industry to pay more attention to social media such as WeChat for its increasingly important role in acquiring medical knowledge and continuing education for Chinese doctors.

The paper "Recognizing Continuous and Discontinuous Adverse Drug Reaction Mentions from Social Media Using LSTM-CRF" presented an important use of social media in medicine, particularly concerning the treatment experiences of patients using mobile devices with valuable adverse drug reaction information (ADR). The authors proposed a deep neural network LSTM-CRF by combining LSTM neural networks with CRFs to recognize ADR mentions from social media in medicine. They investigated the effects of three factors including representation for continuous and discontinuous ADR mentions, subject of posts, and external knowledge bases. The authors believe that the paper is the first attempt to use deep neural networks for mining continuous and discontinuous ADRs from social media.

The paper "A Mobile-Based Question-Answering and Early Warning System for Assisting Diabetes Management" showed that developing a convenient device on chronic disease management and monitoring is increasingly important. Focusing on type 2 diabetes, the authors developed a mobile-based diabetes question-answering and early warning system named Dia-AID, which assisted diabetes patients or people with a high risk of diabetes. The system presented a large multilanguage repository of frequently asked diabetes questions, a multimode fusion Q&A framework, and a health data management module. The system is expected to assist diabetes patients or people at a high risk of diabetes in providing diabetes information and monitoring their health status through diabetes question answering, risk assessment, and health record management.

The paper "The Hierarchies of Multivalued Attribute Domains and Corresponding Applications in Data Mining" described the usage of association rules among attributes as common knowledge patterns in providing potential useful information such as mobile users' interests. The authors concentrated on processing relations between objects and attributes. To reduce the number of association rules for improving computational efficiency, they proposed a method based on concept lattice and attribute analysis and established the connection between the functional dependencies in original relations and their corresponding 'rough' relations. The experiments proved that the method was feasible and effective in the reduction of association rules.

The paper "Automatic Approach of Sentiment Lexicon Generation for Mobile Shopping Reviews" addressed the need of processing large amount of user comments on various products with the increasing use of smartphones. The sentiment analysis of the product reviews heavily relied on the quality of sentiment lexicons. Focusing on the generation of high quality sentiment lexicons, the authors proposed an automatic approach for constructing a domain-specific sentiment lexicon by considering the relationship between sentiment words and product features in the mobile shopping reviews. The generated lexicon was evaluated with a sentiment classification task using product reviews in both English and Chinese. The experiment showed the effectiveness of the proposed sentiment lexicon generation approach in the mobile environment.

The paper "Using Sentence-Level Neural Network Models for Multiple-Choice Reading Comprehension Tasks" studied the multiple-choice task for reading comprehension based on several test datasets. It assessed the usefulness of sentence comprehension versus word comprehension. The authors proposed a sentence-level neural network model using LSTM network and trained a sentence-level attention model to obtain sentence-level representation from sentence embedding in documents. The experiment showed that the model outperforms various state-of-the-art baselines on the multiple-choice reading comprehension task.

The paper "Recommending Mobile Microblog Users via a Tensor Factorization based on User Cluster Approach" explored the factor of user influence for microblog user recommendation in mobile social network. It highlighted the weakness of existing user influence analysis research for ignoring user temporal features and failing to filter marketing users with low influence. The authors proposed a tensor factorization-based user cluster model for influence analysis. This model identified latent influential users and constructed a feature tensor for user influence prediction and microblog user recommendation. The experiment showed the model was able to recognize latent influential users and improve existing recommendation systems through influence analysis.

Conflicts of Interest

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

Acknowledgments

We would like to express our gratitude to all the authors for their high-quality submissions and all the reviewers for their timely and professional reviews.

Tianyong Hao
Raymond Wong
Zhe He
Haoran Xie
Tak-Lam Wong
Fu Lee Wang

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 9434239, 11 pages
https://doi.org/10.1155/2018/9434239

Research Article

Recommending Mobile Microblog Users via a Tensor Factorization Based on User Cluster Approach

Xiangwen Liao,[1,2,3] **Lingying Zhang**,[1,2] **Jingjing Wei**,[4] **Dingda Yang**,[1,2] **and Guolong Chen**[1,2]

[1] *College of Mathematics and Computer Science, Fuzhou University, Fuzhou, China*
[2] *Fujian Provincial Key Laboratory of Network Computing and Intelligent Information Processing, Fuzhou University, Fuzhou, China*
[3] *Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou, China*
[4] *College of Electronics and Information Science, Fujian Jiangxia University, Fuzhou, China*

Correspondence should be addressed to Xiangwen Liao; liaoxw@fzu.edu.cn

Received 27 March 2018; Accepted 31 July 2018; Published 3 October 2018

Academic Editor: Tianyong Hao

User influence is a very important factor for microblog user recommendation in mobile social network. However, most existing user influence analysis works ignore user's temporal features and fail to filter the marketing users with low influence, which limits the performance of recommendation methods. In this paper, a Tensor Factorization based User Cluster (TFUC) model is proposed. We firstly identify latent influential users by neural network clustering. Then, we construct a features tensor according to latent influential user's opinion, activity, and network centrality information. Furthermore, user influences are predicted by the latent factors resulting from the temporal restrained CP decomposition. Finally, we recommend microblog users considering both user influence and content similarity. Our experimental results show that the proposed model significantly improves recommendation performance. Meanwhile, the mean average precision of TFUC outperforms the baselines with 3.4% at least.

1. Introduction

Microblogging services, such as Twitter or Weibo, have been one of the most popular platforms for individuals to exchange information by posting messages or comments in up to 140 characters. With the rapid growth of mobile devices, microblog has created mobile applications to provide their users instant and real-time access from anywhere they can access to the Internet. For example, as of September 2017, the number of monthly active users in Sina Weibo platform is more than 376 million, in which about 92% users are authenticated through mobile phone and/or tablet. A large amount of valuable content exists in the microblog generated data. However, as a result of the rapid increasing population on microblog platform, most users are confronted with the serious problem of information overload [1]. It is extremely difficult to find desirable information using mobile devices. In this situation, recommending relevant users for alleviating the flooding of information appears to be very significant for the users [2].

User influence can provide valuable clue about her preference and thus is indispensable for recommending microblog users in mobile social network [3]. Consequently, incorporating user influence into recommender systems has demonstrated to improve recommendation performance and receives a lot of attention. Li et al. [2] considered social influences and their indirect structural relationships and proposed a Topic-level Social Influence-based mircoblog recommendation model to make user prediction. Jiang et al. [4] proved that users' decisions on information adoption can be affected by individual preference and interpersonal influence and then integrated these two factors to construct a scalable algorithm for online behavior prediction. Chen et al. [5] took advantage of tweet content, user social relations, and explicit features and then proposed a collaborative ranking model for tweet recommendation task. Yan et al. [6] presented a graph-theoretic method to rank tweets and their authors simultaneously by utilizing several networks, i.e., user network, tweet network, and the network that ties the two together. Therefore, it is significant to analyze user

influence in mobile social network and integrate it into the recommendation framework.

There exist several pioneer studies on user influence analysis in mobile microblog platform. Velissarios et al. [7] proposed four different metrics for emphasizing Twitter content features and the behavior of each user's followers and then identified influential users through the comprehensive metrics considering user's affiliation and her interest rate. Mao et al. [8] introduced a learning-based method for analyzing and measuring users' social influence via predicting users' capability of propagating information. Both information extracted from social network structures and user behavior factors were combined in the method to gain a better performance. Xia et al. [9] explained the propagation mechanism of influence in terms of the diffusion of users' emotion. David et al. [10] analyzed the probabilities of one user activated by another user. Then, they combine that user's other features to obtain influence score. Cai et al. [11] proposed an OOLAM model to measure user opinion influence, they separated users interaction graph into two parts, positive graph and negative graph. They ranked users with a PageRank analogous algorithm. These methods reviewed above explored user influence from the perspective of users, which had low accuracy in specific topics.

Recently, various studies are involved in investigating topic level user influence. Those studies showed that most information was created and diffused in terms of topic. User influence can be measured more elaborately from the point of the topic. Therefore, topic level user influence analysis has received increasing attention from researchers. Weng et al. [12] proposed the TwitterRank to calculate user influence score according to the graph structure and topic similarity. Cui et al. [13] introduced item level influence using probabilistic hybrid factor matrix factorization. Chen et al. [14] proposed the MIRC algorithm which can distinguish users in different groups. Their experimental results showed that different influence roles may have stronger influence in their own role level. Wang et al. [15] calculated user influence with four features, i.e., Expert, Leader, Social, and Similar, and then applied user influence to group recommendations. Wei et al. [16] took users' opinion and topic relevance into consideration, and then predicted user influence according to the latent factors resulting from the tensor factorization.

However, most studies for user influence analysis on topic level only consider users' explicit features which can be obtained from users' profile directly [14, 15]. In particular, these existing works neglect the temporal characteristic which can be obtained from the interactions [17]. In addition, the tensor factorization algorithm of user influence analysis tends to give low propagation ability users a high ranking score, since it reduces dimensionality by retaining the critical factors. In this paper, a Tensor Factorization based on User Cluster (TFUC) model [18] is proposed for recommending users according to a specific topic. The TFUC model firstly clusters influential users into a certain groups according to their temporal characteristic. Then, we measure users' influence scores by the temporal restrained CP decomposition on the influential clusters. Finally, both user's influence and content similarity are integrated for recommending users for a given topic. The experimental results in Sina Weibo dataset show that user influence ranking precision of TFUC is better than existing models such as TwitterRank, OOLAM, and HF_CP_ALS. Moreover, our proposed TFUC model can significantly outperform the baseline methods according to recommendation precision.

This is an extension of our previous work [18], in which we proposed a tensor factorization based user influence analysis method. In this paper, we addressed the problem of recommending users in mobile social network and claimed that user influence is a very important factor for user recommendation. We expanded the experiment dataset and added the experiments on recommendation. To summarize, the main contributes of the work are listed as follows.

(1) The latent influence users are identified by a neural network clustering model. This model can filter the marketing users with low influence before constructing tensor, which is proven to significantly enhance the recommendation effect.

(2) TFUC model is proposed by incorporating temporal features, which can improve user recommendation accuracy. Particularly, our method integrates temporal features by tensor model, predicts user influence using temporal restrained CP decomposition, and finally recommends users considering both user influence and content similarity.

(3) We conduct extensive experiments using real-world Tecent Weibo dataset to verify the effectiveness of our proposed recommendation approach. The experimental results suggest that the proposed method can considerably improve recommendation precision and outperform the baseline approaches.

The rest of the paper is organized as follows. The recommendation problem is defined in Section 2. The proposed model for user recommendation is presented in Section 3. Experiments are conducted in Section 4. We conclude the work in Section 5.

2. Problem Formulation

It is well known that people tend to trust a user with high social influence in social network. Therefore, we apply users influence analysis to recommend items for users according to different topic in social media. In this paper, the items refer to users.

We denote items in the microblog as $U_P = \{u_{p_1}, u_{p_2}, u_{p_3}, \ldots, u_{p_n}\}$, where n is the number of items. In the meantime, users who have ever interacted with these items are denoted as $U_c = \{u_{c_1}, u_{c_2}, u_{c_3}, \ldots, u_{c_m}\}$. There are two core elements for the recommendation system: characteristic model of items and characteristic model of users. We characterize every items in U_P as $KW(u_p) = [kw_{p1} : w_{p1}, kw_{p2} : w_{p2}, \ldots, kw_{pn} : w_{pn}]$ and characterize every users in U_c as $KW(u_c) = [kw_{c1} : w_{c1}, kw_{c2} : w_{c2}, \ldots, kw_{cn} : w_{cn}]$.

The goal of this work is to calculate the similarity among items and users and recommend the most similar items for users. In particular, we apply the influence scores of items while calculating the similarity to verify whether the higher influence score the user has, the more likely his

recommendation is to be accepted. To obtain these influence scores, we introduce some necessary features of users, such as the number of fans and the number of posts. Therefore, we let $F = \{f_1, f_2, \ldots, f_m\}$ represent users' fan characteristic and $P = \{P_1, P_2, \ldots, P_m\}$ represent users' post characteristic. Every interaction between U_P and U_c contains the time of when it takes place, so we present this data as $INTERACTION = \{(\text{int}_1, t_1), (\text{int}_2, t_2), \ldots, (\text{int}_q, t_q)\}$.

Referring to Varun's theory [19], we can hardly get user' influence score from single aspect. Thus, we analyze users' influence from four aspects as follows.

(1) Users' propagation ability: getting propagation ability of users is an important purpose in social networks. This ability is usually calculated from the accumulation of time in document collection D. We denote this ability of u_{pi} as $I_{ca}(u_{pi}, U_c, D)$.

(2) Users' opinion strength [20]: one's opinion strength captures his whole tendency and effectiveness in social network. By calculating all of the users' opinion polar who has interacted with u_{pi}, we can get an opinion score of u_{pi}. We present this score as $I_o(u_{pi}, U_c, D)$ which can be analyzed from the document collection D.

(3) Users' fans activity [21]: users with higher levels of activity may contribute more influence to other users in microblog social network. In our work, we regard the number of articles which are posted by u_{cj} as his activity. We can obtain u_{pi}'s global fans activity by accumulating all u_{cj}'s activity who has ever interacted with any of u_{pi}. Formally, we define $I_l(u_{pi}, U_c, P_c)$ as u_{pi} 's global fans activity, where P_c is the collection of users' post features.

(4) Users' network centrality: according to [9, 22], users who have higher influence may have more number of fans. If a user's fans have more fans, it means that the information posted by this user may spread wider. This spreading effect is known as the network centrality and denote as $I_v(u_{pi}, U_c, F_c)$.

Overall, we formalize user influence analysis as follows: given a topic a, the goal is to find a mapping $\text{Inf}_a(U_p, U_c, F_c, P_c, D) \longrightarrow (I_o, I_l, I_v)$. Users influence scores are calculated by aggregating four users' features I_{ca}, I_o, I_l, I_v. After calculating all basic users' influence scores, we can obtain a user influence ranking list sorted by influence scores.

3. Recommendation with User Influence Analysis

In this section, we propose a user influence analysis model [18] and then integrate it into recommendation. A user with high influence can receive a large number of comments in a short time. A user prefers accepting influential users (referred to items) when he is receiving recommended users by recommendation system. Therefore, the performance of recommendation system will be improved by involving the influence of items. Since the factorization based method performs poorly at low ranking users, we design a two-steps method for influence analysis. In the first step, low influence score clusters are identified by a neural network clustering method. In the first step, user influence is predicted by a tensor factorization method.

3.1. Neural Network Clustering Model. Users' global influence consists of multiple individual influence features, i.e., propagation ability, opinion strength, fans activity, and network centrality. The users with higher influence rank would have more comment and stronger opinion strength and are more centrical in the network. On this basis, we first partition data into clusters and filter users with low influence in U_P. We firstly describe how we obtain those four users' features.

(1) Let J denote the number of users who has interacted with u_{pi}. Within a time window t, we can get the delay between the time of u_{pi}'s first interaction happened in D and the time of u_{cj} interacted with u_{pi} according to [23] as follows: assume that the delay $(t_{u_{cj}} - t_{u_{pi}})$ has the exponential distribution form like $f(t_{u_{cj}} \mid t_{u_{pi}}) = \beta_{u_{pi}u_{cj}} \exp(-\beta_{u_{pi}u_{cj}}(t_{u_{cj}} - t_{u_{pi}}))$, where β_{uv} is the transmission rate parameter. The transmission rate parameter captures the capability that how wide a user can reach in the network and thus the computing process is

$$\beta_{u_{pi}u_{cj}}(D) = \frac{\rho_{u_{pi}u_{cj}}(D)}{\Delta_{u_{pi}u_{cj}}(D)} \tag{1}$$

$$\rho_{u_{pi}u_{cj}}(D) = \sum_{di \in D} \delta\left(pu = u_{pi}\right) \delta\left(cu = u_{cj}\right) \tag{2}$$

$$\Delta_{u_{pi}u_{cj}}(D) = \sum_{di \in D} \delta\left(pu = u_{pi}\right) \delta\left(cu = u_{cj}\right) \left(t_{u_{cj}} - t_{u_{pi}}\right) \tag{3}$$

where pu are the basic users and cu are the users who have interacted with pu', the indicator function $\delta(a = b)$ is 1 if $a = b$ is true and 0 otherwise. Equation (2) result in the fact that the total number of times u_{cj} has interacted with u_{pi} and (3) captures a time accumulation of those interactions. After calculating $\beta_{u_{pi}u_{cj}}$, we can infer time accumulation of pu by following aggregate function:

$$I_{ca}\left(u_{pi}\right) = \sum_{j=1}^{J} 1 - \exp\left(-\beta_{u_{pi}u_{cj}}(D)\, t\right) \tag{4}$$

(2) Each user would show an opinion polar to an interbehavior which can be inferred from the interaction between him and basic users. Therefore, we can get u_{pi} 's global opinion strength by accumulating all opinion polar of his interactions. We utilize (5) to obtain the opinion strength of u_{pi}:

$$I_o\left(u_{pi}\right) = \sum_{j=1}^{J} O\left(u_{cj}\right) \tag{5}$$

the indicator function $O(u_{cj})$ is -1 if u_{cj} has ever expressed a negative interbehavior and 1 if u_{cj} has ever expressed a nonnegative interbehavior.

(3) As we defined previously, u_{pi} 's fans activity is related to the total number of articles that are posted by all of his fans.

Based on this definition, we can obtain u_{pi}'s fans activity as follows:

$$I_l\left(u_{pi}\right) = \sum_{j=1}^{J} p_{u_{cj}} \tag{6}$$

(4) Recall that the number of fans of user u_{cj} is available directly from F_c; we calculate user u_{pi}'s network centrality as follows:

$$I_c\left(u_{pi}\right) = \sum_{j=1}^{J} f_{u_{cj}} \tag{7}$$

We now discuss how to partition users in U_P into clusters according to those four influence features. The input samples of our method are U_P. Each sample u_{pi} involves four features which we obtain previously. We denote each sample as $\mathbf{Y} = [y_{i1}, y_{i2}, y_{i3}, y_{i4}]$, where y_{i1}-y_{i4} is $I_{ca}(u_{pi})$, $I_o(u_{pi}), I_l(u_{pi}), I_c(u_{pi})$, respectively. Let C_n denote multiple clustering centers. Each center C_i has four elements, i.e., $[c_{i1}, c_{i2}, c_{i3}, c_{i4}]$. For the clustering problem, the loss function is

$$L\left(X; I_{ca}, I_o, I_l, I_c\right) = \frac{1}{2}\sum_{n,i,j}\left(w_{ij}y_{ni} - c_{ki}\right)^2 \tag{8}$$

where C_k is the clustering center of X and w_{ij} is the weight of between input and interlayer.

We update each w_{ij} using stochastic gradient descent.

$$w_{ij}{}^{(t+1)} = w_{ij}{}^{(t)} - \eta\frac{\partial L}{\partial w_{ij}{}^{(t)}} \tag{9}$$

where

$$\frac{\partial L}{\partial w_{ij}{}^{(t)}} = \frac{\partial\left(\left(w_{ij}{}^{(t)}y_i - c_{ki}\right)^2\right)}{\partial w_{ij}{}^{(t)}} = 2w_{ij}{}^{(t)}y_i{}^2 - 2y_ic_{ki} \tag{10}$$

Bringing (10) into (9), we have

$$w_{ij}{}^{(t+1)} = w_{ij}{}^{(t)} - \eta\left(w_{ij}{}^{(t)}y_i{}^2 - y_ic_{ki}\right) \tag{11}$$

We update clustering centers for each batch as

$$c_{ki} = \frac{\sum_{i,j} w_{ij}\delta_{C_k}\left(y_i\right)}{count_{C_k}\left(Y\right)} \tag{12}$$

where $\delta_{C_k}(x_i)$ is an indicator function for clustering center C_k and the result is 1 if y_i belongs to the cluster C_k and 0 otherwise. The denominator in (12) is a counting function which returns the number of samples in cluster C_k.

3.2. Construction of Tensor User Influence Model. We assign each cluster to a specific influence category. Specifically, the assignment with most of latent influential users in U_P is selected to construct the tensor model. Users in this cluster are denoted as $U_P{}'$, where $U_P{}' \subseteq U_P$. Our users influence model is represented by a 3-order tensor $\mathbf{X} \in R^{I\times J\times K}$, where I is the number of users in $U_P{}'$, J is the number of comment users in U_c, and K is the number of influence features. Tensor decomposition is generally used to predict the distribution of data and the latent features of data. Tensor is used widely in many research area, such as weather forecast, event prediction [24], information recommendation [25], and picture processing [26–28]. Finally, we take these influence features into each tensor slice.

(1) The opinion slice of users: this slice indicates every users' interaction opinion in U_c on $U_P{}'$ in detail; i.e.,

$$X_{ij1} = O\left(u_{cj}\right)\delta\left(cu = u_{cj}\right)\delta\left(pu = u_{pi}\right) \tag{13}$$

where $O(u_{cj})$ is an indicator function as same as the function in (5), $u_{pi} \in U_P{}'$.

(2) The fans activity slice of users: in this slice, users who have ever interacted with u_{pi} would have a activity influence upon u_{pi}. Thus every element in this slice can be represent as

$$X_{ij2} = p_{u_{cj}}\delta\left(cu = u_{cj}\right)\delta\left(pu = u_{pi}\right) \tag{14}$$

(3) The centrality slice of users: as mentioned in Section 3, we present users' network centrality by his diffusion ability which can be presented by his total number of neighbours; i.e.,

$$X_{ij3} = f_{u_{cj}}\delta\left(cu = u_{cj}\right)\delta\left(pu = u_{pi}\right) \tag{15}$$

3.3. Factorization of Tensor User Influence Model. For the tensor $\mathbf{X} \in R^{I\times J\times K}$, the loss function of rank-R CP decomposition [29, 30] is

$$L\left(\mathbf{X}; \mathbf{A}, \mathbf{B}, \mathbf{C}\right) = \sum_{ijk}\left(\mathbf{X}_{ijk} - \sum_{r=1}^{R}\mathbf{A}_{ir}\mathbf{B}_{jr}\mathbf{C}_{kr}\right)^2 \tag{16}$$

The corresponding objective function for stochastic optimization problem is

$$\min_{\lambda,A,B,C} L\left(\mathbf{X}; \mathbf{A}, \mathbf{B}, \mathbf{C}\right) \tag{17}$$

However, temporal influence feature neglects this problem. Thus, a time constraint is added to the user matrix. So the influence score of users whose propagation ability is strong will increase and the score of users who postfrequently receive few comments. The new loss function is written as

$$\begin{aligned} L_\rho = {} & \frac{1}{2}\sum_{ijk}\left(\mathbf{X}_{ijk} - \sum_{r=1}^{R}\mathbf{A}_{ir}\mathbf{B}_{jr}\mathbf{C}_{kr}\right)^2 \\ & + \frac{1}{2}\rho\sum_{i=1}^{I}\sum_{r=1}^{R}\mathbf{Q}_{ii}\left|\mathbf{A}_{ir}\right|^2 + \frac{1}{2}\rho\left(\|\mathbf{B}\|^2 + \|\mathbf{C}\|^2\right) \end{aligned} \tag{18}$$

where $\mathbf{Q}$ is the time constraint matrix which can be obtained from (4). $\mathbf{Q}$ is diagonal and the main diagonal element is

$$\mathbf{Q}_{ii} = I_{ca}\left(u_{pi}\right) \tag{19}$$

where u_{pi} is the users in $U_P{}'$.

The object function is

$$\min_{\lambda,A,B,C} L_\rho (\mathbf{X}; \mathbf{A}, \mathbf{B}, \mathbf{C}) \tag{20}$$

Following the method proposed in [30], the gradient of (18) is

$$\frac{\partial_{L_\rho}}{\partial_{\mathbf{A}}} (\mathbf{X}; \mathbf{A}, \mathbf{B}, \mathbf{C}) = -\mathbf{Y} (\cdot, \mathbf{B}, \mathbf{C}) + \mathbf{AT} (\mathbf{B}, \mathbf{C}) + \rho \mathbf{QA} \tag{21}$$

According to the theory proposed by Acar et al. [31], we can get that

$$\mathbf{T} (\mathbf{B}, \mathbf{C}) = \mathbf{B}^T \mathbf{B} \mathbf{C}^T \mathbf{C} \tag{22}$$

$$\mathbf{Y} (\cdot, \mathbf{B}, \mathbf{C}) = \mathbf{X}_{(1)} \mathbf{Z}_1 \tag{23}$$

where $\mathbf{X}_{(1)}$ is the model-1 unfolding, $\mathbf{Z}_1 = \mathbf{C} \odot \mathbf{B}$, and $\odot$ is the Khatri-Rao product between $\mathbf{C}$ and $\mathbf{B}$. In the same way, we can get $\mathbf{T}(\mathbf{A}, \mathbf{C})$, $\mathbf{T}(\mathbf{A}, \mathbf{B})$, $\mathbf{Y}(\mathbf{A}, \cdot, \mathbf{C})$, and $\mathbf{Y}(\mathbf{A}, \mathbf{B}, \cdot)$.

We can obtain a rule for updating $\mathbf{A}$ by substituting (21) into stochastic gradient descent method as follows:

$$\begin{aligned} \mathbf{A}^{(t+1)} &= \mathbf{A}^{(t)} - \eta^{(t)} \frac{\partial_{L_\rho}}{\partial_{\mathbf{A}}} = \mathbf{A}^{(t)} \\ &\quad - \eta^{(t)} \left[-\mathbf{Y}^{(t)} \left(\cdot, \mathbf{B}^{(t)}, \mathbf{C}^{(t)} \right) + \mathbf{A}^{(t)} \mathbf{T}^{(t)} \left(\mathbf{B}^{(t)}, \mathbf{C}^{(t)} \right) \right. \\ &\quad \left. + \rho \mathbf{Q} \mathbf{A}^{(t)} \right] = \mathbf{A}^{(t)} \left[\mathbf{I} - \eta^{(t)} \mathbf{T}^{(t)} \left(\mathbf{B}^{(t)}, \mathbf{C}^{(t)} \right) \right] \\ &\quad + \eta^{(t)} \mathbf{Y}^{(t)} \left(\cdot, \mathbf{B}^{(t)}, \mathbf{C}^{(t)} \right) - \eta^{(t)} \rho \mathbf{Q} \mathbf{A}^{(t)} \end{aligned} \tag{24}$$

where η is the step size. The updating rules of $\mathbf{B}$ are similar to $\mathbf{C}$. We just give the updating rule of $\mathbf{B}$ due to the space limitation.

$$\begin{aligned} \mathbf{B}^{(t+1)} &= \mathbf{B}^{(t)} - \eta^{(t)} \frac{\partial_{L_\rho}}{\partial_{\mathbf{B}}} \\ &= \mathbf{B}^{(t)} \\ &\quad - \eta^{(t)} \left[-\mathbf{Y}^{(t)} \left(\mathbf{A}^{(t)}, \cdot, \mathbf{C}^{(t)} \right) + \mathbf{B}^{(t)} \mathbf{T}_\rho{}^{(t)} \left(\mathbf{A}^{(t)}, \mathbf{C}^{(t)} \right) \right] \\ &= \mathbf{B}^{(t)} \left[\mathbf{I} - \eta^{(t)} \mathbf{T}_\rho{}^{(t)} \left(\mathbf{A}^{(t)}, \mathbf{C}^{(t)} \right) \right] \\ &\quad + \eta^{(t)} \mathbf{Y}^{(t)} \left(\mathbf{A}^{(t)}, \cdot, \mathbf{C}^{(t)} \right) \end{aligned} \tag{25}$$

3.4. Measurement of Users Influence. We now discuss how to calculate users' influence score by utilizing the result of the tensor decomposition. Users' influence can be calculated from three different influence scores.

(1) Score of users' opinion strength:

$$I_o \left(u_{pi} \right) = \sum_{j=1}^{r} \widehat{\mathbf{X}}_{ij1} \tag{26}$$

(2) Score of users' fans activity:

$$I_l \left(u_{pi} \right) = \sum_{j=1}^{r} \widehat{\mathbf{X}}_{ij2} \tag{27}$$

(3) Score of network centrality:

$$I_v \left(u_{pi} \right) = \sum_{j=1}^{r} \widehat{\mathbf{X}}_{ij3} \tag{28}$$

where $\widehat{\mathbf{X}}$ is the expectation of $\mathbf{X}$. We unify each influence score using min–max normalized method, respectively. And then, we use final influence score by combining these three normalized scores as follows:

$$I \left(u_{pi} \right) = S_i \times \left(I_o \left(u_{pi} \right) + I_l \left(u_{pi} \right) + I_v \left(u_{pi} \right) \right) \tag{29}$$

We add a user topic similarity metric to the combining function to increase users' influence score whose topic similarity is higher. This topic similarity metric will be explained in later sections.

3.5. Recommendation Model with User Social Influence. In this section, we recommend items for users by using content-based recommendation algorithm in Tencent Weibo dataset. The items in this dataset are the person, organization, or group in the real world. Initially, we obtain the preferences and interests of the users whom should receive the recommendation. Users' preferences and interests are analyzed from the articles and comments of them. After that, we establish the users characteristic model based on these preferences and interests. The other essential processes are establishing the items characteristic model. In order to adapt to the dataset, we use the preferences and interests of the items characteristic. Based on these two models, we calculate the similarity between the users and the items. Furthermore, we combine the ranking indicators of items into the similarity and call it influence-similarity. Finally, we recommend items for users according to the influence-similarity.

There are two core parts for the above content-based recommendation process: users characteristic model and items characteristic model. Since the users and the items are all the individual users in Tecent Weibo dataset, we represent every user u to a characteristic vector by using the TF-IDF method. Formally, we denote user vector as $KW(u) = [kw_1 : w_1, kw_2 : w_2, \ldots, kw_n : w_n]$, where kw_i is a word that is extracted from the articles and comments which user u has ever posted, and w_i is the corresponding weight of the word kw_i in the text collection. The weight is calculated by TF-IDF method as follows:

$$TF_{i,j} = \frac{n_{kw_i}}{\sum_k n_{kw_k}} \tag{30}$$

$$IDF = \log \frac{N}{n_i} \tag{31}$$

$$w_i = TF_{i,j} \times IDF \tag{32}$$

where n_{kw_i} is the number of times word kw_i appearing in text j, $\sum_k n_{kw_k}$ is the total number of words that text j contains, N is the total number of texts in the dataset, and n_i is the number of texts that contain word kw_i.

Table 1: Statistics for Sina Weibo dataset.

Topic	Seed users	Articles	Comment Users	Comments
Law	508	6992	8383	14693
Basketball	492	9463	66449	114374
Economy	509	12781	29245	51742
Health	506	9438	30884	54269

Table 2: Statistics Tencent Weibo dataset.

Topic	Item	Received Users	Comments
1.6.2.1	277	23627	44047
1.1.2.1	167	13065	26232
1.2.2.1	147	46550	162939
1.12.4.5	134	15253	27310

The next step is calculate the cosine similarity between the users and items. For example, when calculating the similarity between user u_{c_j} and item u_{r_i}, we have

$$S\left(u_{c_j}, u_{r_i}\right) = \frac{\sum_{i=1}^{M}\left(W_{i,u_{cj}} \times W_{i,u_{ri}}\right)}{\sqrt{\left(\sum_{i=1}^{M} W_{i,u_{cj}}^2\right)\left(\sum_{i=1}^{M} W_{i,u_{ri}}^2\right)}} \tag{33}$$

However, when (33) was adopted to calculate the similarity between user u_{c_j} and item u_{r_i}, it does not take the influence of item u_{r_i} into consideration. Therefore, we add the influence ranking indicator of item u_{r_i} into the original cosine similarity. Thus, the item of higher influence score could have higher probability of being recommended. The cosine similarity with the influence of items is calculate as follows:

$$S\left(u_{c_j}, u_{r_i}\right)' = List\left[I\left(u_{r_i}\right)\right] \times \frac{\sum_{i=1}^{M}\left(W_{i,u_{cj}} \times W_{i,u_{ri}}\right)}{\sqrt{\left(\sum_{i=1}^{M} W_{i,u_{cj}}^2\right)\left(\sum_{i=1}^{M} W_{i,u_{ri}}^2\right)}} \tag{34}$$

where $List[I(u_{r_i})]$ is the influence ranking indicator of item u_{r_i} in its topic areas.

4. Experiments

4.1. Datasets. In this section, extensive experiments are conducted on two popular microblogs' platform in China, i.e., Sina Weibo and Tencent Weibo.

For Sina Weibo dataset, we first crawled 2015 basic users in different topics, including law, basketball, economy, and health. We crawled these users' information and all articles posted by these users from October 31, 2016, to December 1, 2016. The basic statistics of Sina Weibo dataset are showed in Table 1. We annotate users' influence ranking manually according to [16]. In this dataset, the interaction between two users is present as comment. If u_j commented in u_i's articles, there generate an interaction between them. There exists a delay between the time u_i post article and the time u_j commented on this article. Therefore, we can obtain the temporal characteristic based on this delay. Besides, the topic similarity metric in (29) is obtained the same as [16].

For Tencent Weibo dataset, we obtain it from KDD Cup 2012, Track 1. This dataset contains about 6095 high influence users in different topics. These users are called "Item" in this dataset. There are about 73,209,277 recommendation logs in this dataset. The recommendation is send to every user corresponding to a user profile such as his gender, the number of articles he posted, and keywords extracted from all his articles. We can infer user's number of fans from the relational network in this dataset. In this dataset, the interaction between two users is present as recommendation. This interaction contains two significant information, i.e., acceptability and time stamp. Thus, we can obtain u_i's opinion strength and temporal characteristic according to (5) and (3), respectively. For experiment convenience, we choose high influence users in four topics and the time window is from October 12, 2011, to October 13, 2011. Table 2 shows the basic statistics of this dataset. In this dataset, the topic similarity metric in (29) is obtained according to the Jaccard similarity of their characteristic set.

4.2. Baseline. We compared TFUC with the following baselines:

(i) TwitterRank [12], which calculated user influence according to the users' interactions in a certain topic.

(ii) TwitterRank_C, in which we apply TwitterRank to calculated user influence based on the latent influential users cluster which obtained by our cluster model.

(iii) OOLAM [11], which is a PageRank analogous method in which interactions are divided into positive and negative parts so that users' opinion influence is calculated in positive and negative parts, respectively.

(iv) OOLAM_c, in which we use latent influential users to construct positive and negative graph, respectively.

(v) OOLAM_SM, in which the users' topic similarity is taken into consideration in OOLAM.

Table 3: Ranking precision for various methods in four topics.

Method	Law			Method	Basketball		
	P@5	*P*@10	*P*@20		*P*@5	*P*@10	*P*@20
TwitterRank	0.20	0.10	0.10	TwitterRank	0.00	0.00	0.00
TwitterRank_C	0.40	0.30	0.40	TwitterRank_C	0.00	0.00	0.15
OOLAM	0.40	0.40	0.60	OOLAM	0.40	0.40	0.30
OOLAM_C	0.20	0.50	**0.65**	OOLAM_C	0.40	0.50	0.55
OOLAM_SM	0.40	0.50	0.60	OOLAM_SM	0.40	0.50	0.40
OOLAM_SM_C	0.20	0.50	**0.65**	OOLAM_SM_C	0.40	0.50	0.55
HF_CP_ALS	0.40	0.40	**0.65**	HF_CP_ALS	0.40	0.30	0.50
HF_CP_ALS_C	0.00	0.30	0.50	HF_CP_ALS_C	0.60	0.30	0.60
CPSGD	0.40	0.50	0.55	CPSGD	0.40	0.40	0.60
Our method	**0.40**	**0.50**	0.60	Our method	**0.60**	**0.50**	**0.70**
Method	Economy			Method	Health		
	P@5	*P*@10	*P*@20		*P*@5	*P*@10	*P*@20
TwitterRank	0.00	0.00	0.05	TwitterRank	0.00	0.00	0.00
TwitterRank_C	0.00	0.10	0.20	TwitterRank_C	0.00	0.00	0.05
OOLAM	0.20	0.30	0.40	OOLAM	0.40	0.40	0.40
OOLAM_C	0.20	0.30	0.40	OOLAM_C	0.40	0.40	0.55
OOLAM_SM	0.20	0.40	0.50	OOLAM_SM	0.40	0.40	0.40
OOLAM_SM_C	0.20	0.40	0.50	OOLAM_SM_C	0.40	0.30	0.55
HF_CP_ALS	0.20	0.30	0.30	HF_CP_ALS	0.00	0.20	0.40
HF_CP_ALS_C	0.20	0.30	0.30	HF_CP_ALS_C	0.00	0.20	0.40
CPSGD	0.40	0.60	0.50	CPSGD	0.40	0.40	0.45
Our method	**0.40**	**0.60**	**0.50**	Our method	**0.40**	**0.40**	**0.65**

(vi) OOLAM_SM_C, in which the cluster model is added in OOLAM_SM.

(vii) HF_CP_ALS [16], which is a tensor model, in which users' opinion and topic relevance are taken into consideration.

(viii) HF_CP_ALS_C, in which the cluster model is added in HF_CP_ALS.

(ix) CP_SGD, in which low influence users are not filtered when we construct the user's tensor.

Besides, we need to verify whether the performance of recommendation system with influence has a better performance than the recommendation system without; we choose a simple recommendation algorithm, i.e., content-base(BC) algorithm to be the baseline.

4.3. Evaluations. The evaluations include 3 ranking precision evaluations and 2 recommended precision evaluations.

$$P@k = \frac{|A_k \cap B_k|}{k} \tag{35}$$

where A_k is the set of real top-k users and B_k is the predicted set of top-k users.

$$AP = \frac{\sum_{i=1}^{n} |A_i \cap B_i| / i}{n} \tag{36}$$

where i denotes i-th rank and n denotes the number of users.

$$MAP = \frac{\sum_a AP^a}{ca} \tag{37}$$

where a is a certain topic and ca denotes the number of topics.

The two recommended precision evaluations are as follows:

$$AP_r = \frac{1}{\sum_{i=1}^{n} r_i} \sum_{i=1}^{n} \frac{r_i \sum_{j=1}^{i} r_j}{i} \tag{38}$$

$$MAP_r = \frac{\sum_a^{ca} AP_r^a}{ca} \tag{39}$$

where a is a certain topic, ca is the number of topics, and n represents the number of users that need to be recommended. r_i is an accepted index, the value of it is 1 if the i-th user was once accepted when he was recommended to other users and 0 otherwise.

Equation (38) is the average precision in a single topic; it reflects the performance of the model in a single topic. Equation (39) reflects the overall performance of the model in all topics. The higher the MAP_r is, the users of higher influence score are more likely to be accept by other users.

4.4. Precision Results of User Influence Ranking. The $P@k$ of different methods in Sina Weibo dataset is shown in Table 3. The $P@k$ of our method is optimal except for the $P@20$ in law topic. To analyze the experimental results in more detail,

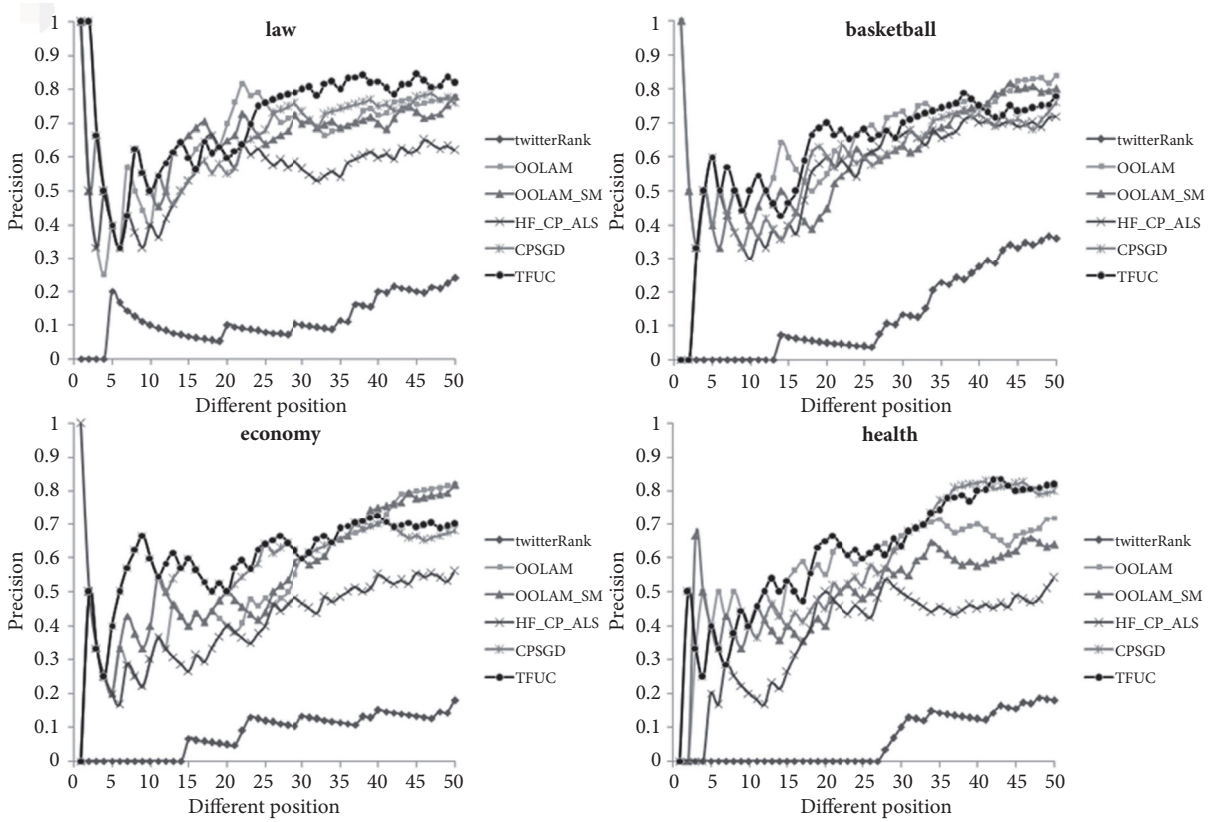


Figure 1: Precision in different position for various methods.

we compare our method with each baseline separately. It can be seen from Table 3 that our proposed method outperforms TwitterRank, which verifies that a user with strong opinion strength, many activity fans, and high propagation ability would be influential. The precision of our method is at least 10% higher than that of OOLAM, which demonstrates that a user with high propagation ability and high topic similarity would has a high influence score. The temporal features are neglected in OOLAM_SM, so that it performs worse than our method. HF_CP_ALS also did not take temporal features into consideration, so that the user with high propagation ability would not get a high influence score. Comparing to CPSGD, the precision of our method has improved at least 10%, which means that filtering some low influential users can improve the performance.

Furthermore, we also calculated the *AP* and *MA*P for each method. Figure 1 shows the precision in different i. The *AP* is higher when the area under the curve is larger. The detail *AP* and *MAP* of each method can be seen in Table 4. *AP* of our method is better than other baselines except for the *AP* of OOLAM in basketball topic and the MAP of our method is best among all methods. We can conclude that our method performs better than other baselines.

4.5. User Influence in Recommendation. In the previous section, we proved that TFUC outperforms other baselines. In this section, we apply TFUC in retrieving users' influence scores in Tencent Weibo dataset. After that, we rank users according to these scores. For each basic users in this dataset, we obtain a recommendation result by counting all u_{pi}'s recommendation logs. If his recommendation was once accepted by other users successfully, his recommendation result could be present as 1 and 0 otherwise. By calculating the correlation coefficient between the influence ranking list and the result list, we could tell which influence analysis method we used in recommendation is closer to the practical situation.

In recommendation task, we also first recognize latent influential users by TFUC model. Now we discuss how to obtain user features to generate clustering model.

(1) As mentioned in datasets description, the interaction between two users is presented as recommendation in Tencent Weibo dataset and this interaction contains time information. Therefore, we can obtain basic users' propagation ability from this time information by (3).

(2) Due to lack of direct opinion information in Tencent Weibo dataset, we present the acceptance of the recommendations as the opinion of a user to the basic users. In this case, we can obtain basic users' opinion strength according to (5).

(3) The users' fans activity and network centrality are obtained similar to (6) and (7).

After getting these four users' features, TFUC model partitions users into different clusters and constructs tensor model based on the users whom in the latent influential

Table 4: AP and MAP comparisons between various methods.

Method	AP				MAP
	Law	Basketball	Economy	Health	
TwitterRank	0.1160	0.1303	0.0818	0.0633	0.0979
TwitterRank_C	0.5139	0.2013	0.1591	0.1267	0.2503
OOLAM	0.6602	0.6570	0.5211	0.5730	0.6028
OOLAM_C	0.6607	**0.6609**	0.5257	0.5572	0.6011
OOLAM_SM	0.6578	0.6027	0.5443	0.4960	0.5752
OOLAM_SM_C	0.6328	0.6418	0.4921	0.5490	0.5790
HF_CP_ALS	0.5577	0.5537	0.4254	0.3695	0.4766
HF_CP_ALS_C	0.4916	0.5271	0.3496	0.3838	0.4380
CPSGD	0.6668	0.5640	0.5828	0.5781	0.5979
Our method	**0.7159**	0.6212	**0.5989**	**0.6110**	**0.6368**

Table 5: The mean precision of the recommendations.

Method	OOLAM	OOLAM_SM	HF_CP_ALS	CPSGD	Our method
MAP_r	0.7028	0.6905	0.6734	0.6964	**0.7104**

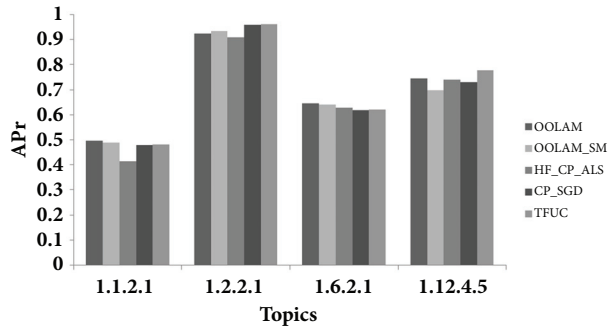


Figure 2: The average precision of the recommendations.

user cluster. After decomposing the tensor model, all users' influence score can be predicted according to (29).

To calculate the precision of the recommendations, we need to obtain the successful accepted list. In Tencent Weibo dataset, if a u_j accepted the recommendation of whose entity is u_i, the system will record this interaction into recommendation logs. By calculating which user is successfully accepted by other users in logs file, we can get an acceptance list.

Finally, we can get a recommendations precision by compare the influence list and the acceptance list according to (39). Figure 2 and Table 5 show the results.

It can be seen from Figure 2 and Table 5 that the recommend results which combine with the users influence ranking list obtained from our method have the similar performance to the OOLAM method in each topic. However, our method shows a better overall performance in four topics. This result illustrates that when the influence scores which add temporal characteristic and topic similarity were applied in the recommendation system, the items with a higher influence are more likely to be accept. The AP_r value of our method has promoted 2% to 8% than OOLAM_SM. The result above reflects that when the temporal characteristic is considered, the items influence ranking list is better adapted to the actual recommendation results. Our method has a higher recommendation precision than the method HF_CP_ALS which also confirmed the above conclusion. Compared with CP_SGD methods, the AP_r value of our method has improved in every topic. This is due to the fact that we filter the impact of low influence but high activity marketing items of which the recommendations have a low probability to be accepted.

Based on the analysis above, we can conclude that the high influence items obtained from our method have a wider range of probability to be accepted by users. Therefore, we combine the recommendation system with the item influence obtained from our method. Firstly, we calculate the recommendations result list from rec_log_train dataset in topics "1.6.2.1", "1.1.2.1", "1.2.2.1", and "1.12.4.5". We next calculate the influence-similarity between the users and items in this recommendations result list. Then, we choose the top 100 most similar results as our recommendations for users and calculate the average recommended precision in each topic. Finally, we obtain the MAP_r by fusing the average recommended precision of each topic. The MAP_r of content-based recommendation method is 14.5. The MAP_r improves to 15.5 when TFUC is integrated. This indicates that the performance of the recommendation system can be improved when the influence is considered.

5. Conclusion

This paper focuses a recommendation task in which users' influence analysis is involved in microblogs. We introduce a two-steps method for influence analysis. Firstly, users are partitioned into influential part and uninfluential part. And then, we expect CP decomposition with stochastic gradient descent method to expedite decomposition. In addition, a time constraint matrix is also involved in the user factor matrix during the decomposition. Finally, we apply TFUC model to recommend items for users according to

the influence of items. The experimental results show that TFUC outperforms other baselines with 3.4% at least. The extensive experiments in Tencent Weibo dataset show that the precision of the recommendation system is improved when we combine the recommendation system with the influence of items.

There are many potential future directions of this work. First, the temporal characteristic of users or items in this paper is estimate rough. The reason is that we suppose that the delay of the interaction satisfies the exponential distribution. We do not conduct in-depth research of the other temporal accumulation models. Such ambiguity can be further aggravated in the microblog. Additionally, the recommendation algorithm of our work is still primitive with challenges including how to design a more realistic recommendation algorithm and combine it with user influence aspect.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research project was supported by the National Natural Science Foundation of China (no. 61772135 and no. U1605251), the Open Project of Key Laboratory of Network Data Science & Technology of Chinese Academy of Sciences (no. CASNDST201606 and no. CASNDST201708), and the Directors Project Fund of Key Laboratory of Trustworthy Distributed Computing and Service (BUPT) Ministry of Education (no. 2017KF01). The authors thank Lin Gui and Kam-Fai Wong for their cooperation in TFUC model.

References

[1] M. G. Rodriguez, K. Gummadi, and B. Schoelkopf, "Quantifying information overload in social media and its impact on social contagions," in *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, pp. 170–179, Ann Arbor, Michigan, USA, 2015.

[2] D. Li, Z. Luo, Y. Ding et al., "User-level microblogging recommendation incorporating social influence," *Journal of the Association for Information Science and Technology*, vol. 68, no. 3, pp. 553–568, 2017.

[3] X. D. Wu, Y. Li, and L. Li, "Influence analysis of online social networks," *Chinese Journal of Computers. Jisuanji Xuebao*, vol. 37, no. 4, pp. 735–752, 2014.

[4] M. Jiang, P. Cui, F. Wang, W. Zhu, and S. Yang, "Scalable recommendation with social contextual information," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 11, pp. 2789–2802, 2014.

[5] K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, and Y. Yu, "Collaborative personalized tweet recommendation," in *Proceedings of the 35th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 661–670, USA, August 2012.

[6] R. Yan, M. Lapata, and X. Li, "Tweet recommendation with graph co-ranking," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL '12)*, pp. 516–525, July 2012.

[7] V. Zamparas, A. Kanavos, and C. Makris, "Real Time Analytics for Measuring User Influence on Twitter," in *Proceedings of the 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 591–597, Vietri sul Mare, Italy, November 2015.

[8] J.-X. Mao, Y.-Q. Liu, M. Zhang, and S.-P. Ma, "Social influence analysis for micro-blog user based on user behavior," *Jisuanji Xuebao/Chinese Journal of Computers*, vol. 37, no. 4, pp. 791–800, 2014.

[9] X. Hu, L. Tang, J. Tang, and H. Liu, "Exploiting social relations for sentiment analysis in microblogging," in *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM '13)*, pp. 537–546, Rome, Italy, February 2013.

[10] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 137–146, Washington, DC, USA, 2003.

[11] K. Cai, S. Bao, Z. Yang et al., "OOLAM: An opinion oriented link analysis model for influence persona discovery," in *Proceedings of the 4th ACM International Conference on Web Search and Data Mining, WSDM 2011*, pp. 645–654, China, February 2011.

[12] J. Weng, E. Lim, J. Jiang, and Q. He, "TwitterRank: finding topic-sensitive influential twitterers," in *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM '10)*, pp. 261–270, New York, NY, USA, February 2010.

[13] P. Cui, W. Fei, S. Yang, and L. Sun, "Shiqiang Yang, and Lifeng Sun. Item-level social influence prediction with probabilistic hybrid factor matrix factorization," in *Proceedings of the In Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI, 2011*, San Francisco , California, USA, 2011.

[14] C. Chen, D. Gao, W. Li, and Y. Hou, "Inferring topic-dependent influence roles of Twitter users," in *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2014*, pp. 1203–1206, Australia, July 2014.

[15] J. Wang, Z. Liu, and H. Zhao, "Topic oriented user influence analysis in social networks," in *Proceedings of the 2015 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology Workshops, WI-IAT Workshops 2015*, pp. 123–126, Singapore, December 2015.

[16] J.-J. Wei, C. Chen, X.-W. Liao, G.-L. Chen, and X.-Q. Cheng, "User social influence analysis based on constrained nonnegative tensor factorization," *Tongxin Xuebao/Journal on Communication*, vol. 37, no. 6, pp. 154–162, 2016.

[17] Y. Zhou and L. Liu, "Social influence based clustering of heterogeneous information networks," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 338–346, ACM, Chicago, Ill, USA, August 2013.

[18] X. Liao, L. Zhang, L. Gui, K. Wong, and G. Chen, "A Tensor Factorization Based User Influence Analysis Method with Clustering and Temporal Constraint," in *Natural Language Processing and Chinese Computing*, vol. 10619 of *Lecture Notes in Computer Science*, pp. 877–886, Springer International Publishing, Cham, 2018.

[19] V. R. Embar, I. Bhattacharya, V. Pandit, and R. Vaculin, "Online topic-based social influence analysis for the Wimbledon championships," in *Proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2015*, pp. 1759–1768, Australia, August 2015.

[20] "Social Media Processing," in *Proceedings of Third National Conference of the Social Media Processing , SMP '14*, H. Huang, T. Liu, H.-P. Zhang, and J. Tang, Eds., vol. 489, Springer, Beijing , China, 2014.

[21] "Proceedings of 24th Annual International Conference on Computer Science and Software Engineering, CASCON, '14," in *Proceedings of 24th Annual International Conference on Computer Science and Software Engineering, CASCON, 2014*, J. Ng, J. Li, and K. Wong, Eds., IBM / ACM, Markham , Ontario, Canada, 2014.

[22] D. Xu, Y. Liu, M. Zhang, and S. Ma, "Study on user influence in online social networks," *Journal of Chinese Information Processing*, vol. 30, no. 2, pp. 83–89, 2016.

[23] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD-2010*, pp. 1019–1028, USA, July 2010.

[24] T. M. Do, "Non-linear Time-series Analysis of Social Influence," in *Proceedings of the the 2016*, pp. 12–16, San Francisco, California, USA, June 2016.

[25] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos, "Tag recommendations based on tensor dimensionality reduction," in *Proceedings of the 2008 2nd ACM International Conference on Recommender Systems, RecSys'08*, pp. 43–50, Switzerland, October 2008.

[26] C. Wang, X. He, J. Bu, Z. Chen, C. Chen, and Z. Guan, "Image representation using Laplacian regularized nonnegative tensor factorization," *Pattern Recognition*, vol. 44, no. 10-11, pp. 2516–2526, 2011.

[27] A. Karami, M. Yazdi, and A. Zolghadre Asli, "Hyperspectral image compression based on tucker decomposition and discrete cosine transform," in *Proceedings of the 2nd International Conference on Image Processing Theory, Tools and Applications, IPTA '10*, pp. 122–125, France, July 2010.

[28] A. Karami, M. Yazdi, and A. Z. Asli, "Best rank-r tensor selection using Genetic Algorithm for better noise reduction and compression of Hyperspectral images," in *Proceedings of the Fifth International Conference on Digital Information Management (ICDIM '10)*, pp. 169–173, Thunder Bay, ON, Canada, July 2010.

[29] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.

[30] T. Maehara, K. Hayashi, and K.-I. Kawarabayashi, "Expected tensor decomposition with stochastic gradient descent," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence, AAAI 2016*, pp. 1919–1925, USA, February 2016.

[31] E. Acar, D. M. Dunlavy, and T. G. Kolda, "A scalable optimization approach for fitting canonical tensor decompositions," *Journal of Chemometrics*, vol. 25, no. 2, pp. 67–86, 2011.

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 1789121, 11 pages
https://doi.org/10.1155/2018/1789121

Research Article

The Hierarchies of Multivalued Attribute Domains and Corresponding Applications in Data Mining

Yuxia Lei,[1] **Yushu Yan**,[1] **Yonghua Han**,[1] **and Feng Jiang**[2]

[1]*School of Information Science and Engineering, Qufu Normal University, Rizhao 276826, China*
[2]*College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China*

Correspondence should be addressed to Yuxia Lei; yx_lei@126.com

Received 30 March 2018; Accepted 4 July 2018; Published 28 August 2018

Academic Editor: Tianyong Hao

In mobile computing, machine learning models for natural language processing (NLP) have become one of the most attractive focus areas in research. Association rules among attributes are common knowledge patterns, which can often provide potential and useful information such as mobile users' interests. Actually, almost each attribute is associated with a hierarchy of the domain. Given an relation R = (U, A) and any cut α_a on the hierarchy for every attribute a, there is another rough relation R_Φ, where $\Phi = (\alpha_a : a \in A)$. This paper will establish the connection between the functional dependencies in R and R_Φ, propose the method for extracting reducts in R_Φ, and demonstrate the implementation of proposed method on an application in data mining of association rules. The method for acquiring association rules consists of the following three steps: (1) translating natural texts into relations, by NLP; (2) translating relations into rough ones, by attributes analysis or fuzzy k-means (FKM) clustering; and (3) extracting association rules from concept lattices, by formal concept analysis (FCA). Our experimental results show that the proposed methods, which can be applied directly to regular mobile data such as healthcare data, improved quality, and relevance of rules.

1. Introduction

With the rapid growth in use of mobile devices, more and more mobile generated data is in great need of processing. A large amount of valuable content exists in natural text such as web pages, news feeds, and Twitter/WeChat messages. Natural language processing (NLP) techniques have proven to be useful in dealing with the information overload problem in the mobile environment, for example, news summarization, question answering, and information extraction and retrieval. In these areas, machine learning models for NLP are one of the important research contents [1], in which association rules are common knowledge patterns. Association rules can often provide potential and useful information for mobile clients, contributing to automatically providing personalized recommended services.

In the process of knowledge acquisition from natural texts, we frequently encounter multivalued attributes such as spatial locations and security policies in mobile environments. Actually, for each multivalued attribute, there are different levels of partitions or fuzzy partitions in its domain, which are forbidden in the domain. If we take the different partitions as new values and then obtain attribute values of different granularity. It will be meaningful to discover the attribute dependence of different granularity.

Formal concept analysis (FCA) is an effective tool for knowledge representation, acquisition, and knowledge discovery. FCA focuses on the concept lattice induced by a binary relation between a pair of sets (called objects and attributes, respectively). A node of concept lattices is an objects/attributes pair, called a (formal) concept, consisting of two parts: the extent (objects the concept covers) and intent (attributes describing the concept). The line diagram corresponding to a concept lattice vividly unfolds generalization/specialization relationship among concepts [2]. Recently, concept lattices have already been successfully applied to a wide range of scientific disciplines including knowledge representation [3–5], knowledge discovery [6–8], knowledge reduction [9–11], hybrid relation analysis [12], wireless sensor network [13], and information retrieval [14].

In order to process natural texts, we firstly extract some formal objects and attributes using NLP, secondly translate the texts into relations, and thirdly process the relations using FCA. However, with the growth of the size of the relation, the number of concepts and association rules grows in an exponential manner. Hence, it is necessary to reduce contexts before applying FCA. Several researchers have used matrix approximation techniques such as singular value decomposition (SVD) [15] and nonnegative matrix factorization (NMF) [11] for reducing the size of the context. However, matrix DR methods are based on the expensive eigenvalue computations and hence are known for their high computational complexity. Cluster analysis can well be used as method for data reduction under the notion of concept decomposition (CD). With its lesser computational complexity, FKM [15] is proved to be an alternative method for reducing the dimensionality of context, thereby controlling the size of concept lattices. Actually, there are different levels of partitions or fuzzy partitions [16, 17] in each multivalued attribute domain. Therefore, each attribute is associated with a hierarchy of the domain of the attribute. Based on attribute analysis and FKM, this paper proposes a method of acquiring association rules, which can validly reduce the number of association rules. The method firstly translates attribute values into different abstract hierarchies, which are new attribute values, and then generates association rules with FCA. Experimental results on heart disease data show that the proposed method can improve quality and relevance of association rules, which can be applied directly to regular mobile data such as healthcare data and spatial locations, contributing to providing precise personalized recommended services.

The paper is organized as follows: the next section gives the basic notations of FCA and FKM; the third section gives the hierarchies of domains, discusses the possible connection between the functional dependencies of R and R_Φ, and gives a sufficient and necessary condition for a functional dependence $B \longrightarrow C$ holding in both R_Φ and R_Ψ if Ψ is below Φ; the fourth section provides some experiments, in order to verify the feasibility of the method. Finally, Section 5 concludes the paper.

2. Some Basic Notations

2.1. Translation from Natural Texts into Relations. Before using FCA to extract association rules, we need to translate natural texts into a knowledge frame and then merge the related frames into a relation. Our method can be described as follows: firstly, we create some attribute thesaurus. For example, we create an attribute thesaurus that describes people such as age, sex, height, weight, date of birth, hobbies, occupation, etc. Secondly, using a semiautomatic approach, we translate text knowledge into a frame and then merge some related frames into a relation.

2.2. Formal Concept Analysis. This section will introduce some basic notations in FCA [2]. A (formal) context is defined as a triple $K = (G, M, I)$, where G and M are sets and $I \subseteq G \times M$ is a binary relation. For any $X \subseteq G$ and $Y \subseteq M$, the pair (X, Y) is called a (formal) concept if (1) Y is the set of attributes common to the objects in X and (2) X is the set of objects having all attributes in Y. X and Y are called the (concept) extent and the (concept) intent of the concept, respectively. There are two kinds of special concepts: object concepts and property concepts. Given an object $g \in G$, the object concept of g is the smallest concept having g in its extent. Correspondingly, given an attribute $m \in M$, the attribute concept of m is the greatest concept having m in its intent.

The line diagram corresponding to a concept lattice can vividly unfold generalization-specialization relationship among concepts. The labeling can be simplified considerably by putting down each object and each attribute only once. Thus, the concept lattices can be described by the line diagrams with reduced labeling. In a line diagram, the name of an object g is always attached to the circle representing the smallest concept with g in its extent; dually the name of an attribute m is always attached to the circle representing the largest concept with m in its intent. This allows us to read the map I from the diagram: an object g has an attribute m if and only if there is an ascending path from the circle labeled by g to the circle labeled by m. The extent of a concept consists of all tuples whose labels are below in the diagram and the intent consists of all properties attached to concepts above in the hierarchy. Thus, we can easily extract association rules with 100% confidence from the line diagrams, and the *stem base* of the attribute implications is nonredundant and complete [2].

As many practical applications involve nonbinary data, multivalued contexts have been introduced in FCA. A multivalued context $K = (G, M, W, I)$ consists of sets G, M, W and a ternary relation I between G, M, and W for which it holds that $(g, m, w) \in I$ and $(g, m, v) \in I$ always imply $w = v$. The elements of G, M, and W are called objects, attributes, and attribute values, respectively. A tuple (g, m, w) is interpreted as object g that has value w for attribute m. Actually, a multivalued context can be regarded as a relation with the column containing the objects being a primary key. In the RDM, a relation is described by a relation schema $S = R(m_1, m_2, \ldots, m_n)$, where m_i $(1 \le i \le n)$ represent attributes. Each attribute m is associated with a domain D_m, which is the set of possible values for the attribute m. A relation (R, A) is denoted by a set of tuples $U = \{r: r = (v_1, v_2, \ldots, v_n) \in R\}$, where $A = (m_1, m_2, \ldots, m_n)$ and g is a tuple such that for every $1 \le i \le n$, $v_i \in D_{mi}$. An equivalent way to view such a tuple $g \in U$ is as a map from A to $\prod_{a \in A} D_m$ such that $g(m) \in D_m$ [11]. Thus we can further represent a relation R by a triple (U, A, I), where I is a map from $U \times A$ to $\bigcup_{a \in A} D_m$ such that, for any $(g, m) \in U \times A$, $I(g, m) = g(m) \in D_m$. A relation (U, A, I) can be thought of as representing a table with rows corresponding to U, columns corresponding to A, and table entries at the intersection of rows and columns containing values in domains.

2.3. Fuzzy K-Means Clustering. Fuzzy k-means (FKM) [18] partitions a set of t-dimensional vectors $X = \{X_1, X_2, \ldots, X_n\}$ into k-clusters, where $X_j = \{X_{1j}, X_{2j}, \ldots, X_{tj})$ represents the jth sample. For X_j and the ith cluster center v_i, there is a membership degree u_{ji} indicating to what degree sample X_j belongs to v_i, $i = 1, 2, \ldots, k$. Thus, there is a fuzzy

Table 1

Category	Systolic(mmHg)	Diastolic(mmHg)
Normal value	120	80
Normal high value	120 ~139	80 ~89
High blood pressure	140	90
(Hypertension) stage 1	140 ~159	90 ~99
(Hypertension) stage 2	160 ~179	100 ~109
(Hypertension) stage 3	180	110
Simple systolic hypertension	140	89

partition matrix $U = (u_{ij})_{d\times k}$. The FKM algorithm is based on minimizing the objective function J_{fuzz} defined as

$$J_{fuzz} = \sum_{j=1}^{d}\sum_{i=1}^{k} {u_j}^m {d_i}^2 \quad (1)$$

where d_j is the Euclidean distance between X_j to the cluster center v_i. The exponent m in (1) is called fuzzifier parameter and it defines the fuzziness of the clustering. The formulae of u_{ji} and v_i are

$$u_{ji} = \frac{1}{\sum_{p=1}^{k}\left(d_j/d_p\right)^{1/(m-1)}}, \qquad v_i = \frac{\sum_{j=1}^{d} {u_{ji}}^m X_j}{\sum_{j=1}^{d} {u_{ji}}^m} \quad (2)$$

where $m \neq 1$ and $i = 1, \ldots, k$.

Based on the above discussion, the FKM algorithm can be summarized as follows.

Step 1. Choose the number of clusters k, degree of fuzziness m, and a threshold value e. Initialize the fuzzy partition matrix U.

Step 2. Compute the cluster centers v_i ($i = 1, 2, \ldots, k$), according to (2).

Step 3. Compute the Euclidean distance d_{ji} from the sample X_j to the cluster center v_i according to the Euclidean distance. Then calculate all uji using (2) and update fuzzy partition matrix U.

Step 4. Compute the objective function J_{fuzz} using (1). Verify whether the function converges or the difference between the two adjacent values of objective function is less than the given threshold value e, then stop. Otherwise repeat from the Step 2.

3. Attribute Analysis and the Reducts in Rough Relations

3.1. Attribute Analysis. For attributes, there are often different criteria for division, for example, Chinese blood pressure categories and Chinese age categories (≥18 years old), which can be found in "Guidelines for the Prevention and Control of Hypertension in China" (2005 Revision).

Table 2

Category	Sub-category	Age range
Youth	Puberty	18-28
	Mature period	29-40
Middle aged	Strong period	41-48
	Robust period	49-55
	Adjustment period	56-65
Old age	Initial old period	66-72
	Middle old period	73-84
	Old period	85

Chinese blood pressure categories are described as shown in Table 1.

Chinese age categories are described as shown in Table 2.

Generally, attributes can be divided into the three types. Type 1: there is a category; type 2: there is no category, but there are reference criteria for the classification of attributes; type 3: attribute values are never category or classification criteria for reference. In Algorithm 1, the subscript of R in step 3 is not obvious. Therefore, we followed the methods of Lei et al. 2016 [19] and propose a method for extracting association rules, which has the following steps: (1) analyzing attribute types and the structures of domains, (2) generating different hierarchies of attribute values with FKM clustering, (3) translating original relations into rough ones, (4) generating the concept lattice, and (5) extracting the association rules from concept lattices. The method can be described as in Algorithm 1.

3.2. The Reducts in Rough Relations. In a rough relation in the rough relation databases, each attribute $a \in A$ is associated with a equivalence relation θ_a on domain D_a. We denote the corresponding partition of θ_a on D_a by P_a: ${X^a}_1, {X^a}_2, \ldots, {X^a}_k$ for some natural number k.

Definition 1 (a rough relation). A rough relation R is a subset of $\prod_{a\in A} P(D_a)$ such that, for every $x \in R$, every $a \in A$, and every $1 \leq i \leq k$, $|x(a) \cap {X^a}_i| \neq 1$, where $P(D_a)$ is the power set of D_a.

Hence, in a rough relation R, a tuple x takes multivalues at attributes, satisfying certain conditions, where the conditions are given in terms of equivalence relations on domains of

Algorithm for extracting associate rules from a relation
Input: a relation R = (U, A)
Output: association rules satisfying given the minimum support and the minimum confidence
Process
Step 1: For any attribute a ∈ A, generating a hierarchy H_a of its domain by traditional standards or FKM clustering method.
Step 2: Given a cut $\alpha_a = V_{1a} \cup V_{2a} \cup \ldots \cup V_{ka}$ in H_a, then there is a set $\Phi = (\alpha_a: a \in A)$ of cuts,
translating the relation R into a rough relation R ., where $r(a) = V_{ia}$ in R . if $r(a) \in V_{ia}$ in R.
Step 3: Translating R . into a binary relation R′,
Step 4: Generating the concept lattice from R′,
Step 5: Extracting association rules satisfying given the minimum support and the minimum confidence.

Algorithm 1: Extracting association rules from a relation.

attributes. A rough relation R is reduced to be a normal relation if every attribute has $P(D_a)$ as the domain, instead of D_a.

Given an attribute a, there is a hierarchy $T_a = (S_a, \subseteq)$, where S_a is a set of subsets of D_a and $\subseteq$ is a binary relation on S_a, such that (1) $D_a \in S_a$; (2) for any $v \in D_a$, $\{v\} \in S_a$; (3) $(S_a, \subseteq)$ is a tree.

Definition 2. A cut α is a subset of S_a such that for any path σ from the root to a leaf, $|\alpha \cap \sigma| = 1$; given two cuts α and β, we say that α is above β, denoted by $\alpha \gg \beta$, for any path σ from the root to a leaf, $S_{\beta,\sigma} \subseteq S_{\alpha,\sigma}$, where $S_{\alpha,\sigma}$ and $S_{\beta,\sigma}$ are the unique ones in $\alpha \cap \sigma$ and $\beta \cap \sigma$, respectively.

Given a cut α on T_a, there is an equivalence relation $\theta_{\alpha,a}$ on D_a such that, for any u, $v \in D_a$, $u\theta_{\alpha,a}v$ iff there is a unique $s \in \alpha$ such that u, $v \in s$. We use $[u]_{a,\alpha}$ to denote the equivalence class of $\theta_{\alpha,a}$ containing u.

Let Φ be a cut vector $(\alpha_a : a \in A)$, where α_a is a cut on T_a. Given two cut vectors $\Phi = (\alpha_a : a \in A)$ and $\Psi = (\beta_a : a \in A)$, we say that Φ is above Ψ, denoted by $\Phi \gg \Psi$, if for every $a \in A$, $\alpha_a \gg \beta_a$. Given a relation R and a cut vector $\Phi = (\alpha_a : a \in A)$, there is a relation R_Φ such that for any tuple $x \in R$ and attribute $a \in A$, if $x(a) = u$ in R then $x(a) = [u]_{a,\alpha}$ in R_Φ. Given a relation R and a cut vector Φ, define a relation $\theta_{R,\Phi}$ such that, for any x, $y \in R$, $x\theta_{R,\Phi}y$ iff $x(a) = [u]_{a,\alpha}$ and $y(a) = [v]_{a,\alpha}$ in R_Φ and $[u]_{a,\alpha} = [v]_{a,\alpha}$, where $x(a) = u$ and $y(a) = v$ in R.

Proposition 3. *Given a relation R and a cut vector Φ, $\theta_{R,\Phi}$ is an equivalence relation on U.*

Proposition 4. *Given a relation R and two cut vectors $\Phi = (\alpha_a : a \in A)$ and $\Psi = (\beta_a : a \in A)$, if $\Phi \gg \Psi$ then $\theta_{R,\Psi}$ is a refinement of $\theta_{R,\Phi}$.*

Proof. For any x, $y \in R$, assume that $x\theta_{R,\Psi}y$. Then, for any $a \in A$, $[x(a)]_{a,\beta_a} = [y(x)]_{a,\beta_a}$. Because β_a is a refinement of α_a, $[x(a)]_{a,\alpha_a} = [y(a)]_{a,\alpha_a}$. That is, for any $a \in A$, $[x(a)]_{a,\alpha_a} = [y(a)]_{a,\alpha_a}$, i.e., $x\theta_{R,\Phi}y$. □

Definition 5. Given a relation R and subsets $B, C \subseteq A$, if, for any x, $y \in R$, if $x(a) = y(a)$ for every $a \in B$ then $x(c) = y(c)$ for every $c \in C$, we say that C depends on B in R, denoted by $R| = B \longrightarrow C$.

Assume that $R| = B \longrightarrow C$. By the sense of functional dependencies, we define a function $f : \prod_{b\in B} D_b \longrightarrow \prod_{c\in C} D_c$ such that, for any $v \in \prod_{b\in B} D_b$, if there is a $x \in R$ such that $x(b) = v(b)$ for every $b \in B$ then $f(v) = u \in \prod_{c\in C} D_c$, where u is defined as follows: for any $c \in C$, $u(c) = x(c)$.

By the assumption that $R| = B \longrightarrow C$, f is a function. We say that f witnesses that $R| = B \longrightarrow C$. There are two special cut vectors $\Phi_\perp$ and Φ_T defined as follows: for any $a \in A$, $\Phi_{\perp,a} = \{u: u \in D_a\}$, $\Phi_{T,a} = \{D_a\}$. It is clear that for any cut vector Φ, $\Phi_T \gg \Phi \gg \Phi_\perp$.

Proposition 6. *(i) $R_{\Phi_T,a} = \{D_{a1}, D_{a2}, \ldots, D_{an}\}$. Therefore, for any $a \in A$, $\{a\}$ is a reduct of R_{Φ_T}; (ii) $R_{\Phi_T,a} = R$. Therefore, $B \subseteq C$ is a reduct of $R_{\Phi\perp}$ iff B is a reduct of R.*

Given a relation R and two cut vectors Φ and Ψ, if $\Phi \gg \Psi$ then, for any subsets $B, C \subseteq A$, $R_\Psi| = B \longrightarrow C$ and $R_\Phi| = B \longrightarrow C$ are not related; i.e., it is possible that $R_\Psi| = B \longrightarrow C$ and $R_\Phi| \neq B \longrightarrow C$; or $R_\Psi| \neq B \longrightarrow C$ and $R_\Phi| = B \longrightarrow C$. By Propositions 3 and 4, we have that there are R, Φ, and Ψ such that $\Phi \gg \Psi$, $R_\Psi| \neq B \longrightarrow C$, and $R_\Phi| = B \longrightarrow C$. Let $\Phi = \Phi_T$, such that $\Phi \gg \Psi$ and $R_\Psi| \neq B \longrightarrow C$. Because, for any B, $C \subseteq A$, $R_{\Phi_T}| = B \longrightarrow C$. We give the following example to show that there are R, Φ, and Ψ such that $\Phi \gg \Psi$, $R_\Psi| = B \longrightarrow C$, and $R_\Phi| \neq B \longrightarrow C$.

Example 7. Let $A = \{a_1, a_2\}$, $D_{a1} = \{1, 2, 3, 4, 5, 6\}$, and $D_{a2} = \{1, 2\}$. Let $R = \{x_1, \ldots, x_6\}$, which is described in Table hierarchies(a), Let $\Psi = (\{\{1, 2\}, \{3, 4\}, \{5, 6\}\}, \{\{1\}, \{2\}\})$, and $\Phi = (\{\{1, 2, 3, 4\}, \{5, 6\}\}, \{\{1\}, \{2\}\})$. Then, R_Ψ and R_Φ are represented in Tables 1(b) and 3(c).

We have that $R_\Psi| = a_1 \longrightarrow a_2$ and $R_\Phi| \neq a_1 \longrightarrow a_2$.

Definition 8. Given a relation R and two cut vectors Φ and Ψ, assume that $\Phi \gg \Psi$. Let $b, c \in A$ such that $R_\Psi| = b \longrightarrow c$,

Table 3

(a)

tuples	a_1	a_2
x_1	1	1
x_2	2	1
x_3	3	2
x_4	4	2
x_5	5	1
x_6	6	1

(b)

tuples	a_1	a_2
x_1	[1]	1
x_2	[1]	1
x_3	[3]	2
x_4	[3]	2
x_5	[5]	1
x_6	[5]	1

(c)

tuples	a_1	a_2
x_1	[1]	1
x_2	[1]	1
x_3	[1]	2
x_4	[1]	2
x_5	[5]	1
x_6	[5]	1

and f: $D_{b,\alpha_b} \longrightarrow D_{c,\alpha_c}$ witnesses that $R_\Psi|$ = b $\longrightarrow$ c, where $D_{b,\alpha_b} = \{[u]_{\alpha_b} : u \in D_b\}$.

We say that $R_\Psi|$ = b $\longrightarrow$ c is compatible with Φ in R if for any x $\in$ R, there is a v $\in$ D_c such that $\{[f([u]_\Psi)]_\Phi : u \in [x(b)]_\Phi\} = [v]_\Phi$. Let B, C $\subseteq$ A such that $R_\Psi|$ = b $\longrightarrow$ c, and $f : \prod_{b\in B} D_{b,\alpha_b} \longrightarrow \prod_{c\in C} D_{c,\alpha_c}$ witness that $R_\Psi|$ = b $\longrightarrow$ c. We say that $R_\Psi|$ = b $\longrightarrow$ c is compatible with Φ in R if, for any x $\in$ R, there is a vector ($v_c \in D_c$: c $\in$ C) such that

$$\left\{\left[f\left(\prod_{b\in B}[u_b]_\Psi\right)\right]_\Phi : u \in [x(b)]_\Phi, b \in B\right\} = \prod_{c\in C}[v_c]_{\alpha_c} \quad (*)$$

Proposition 9. *Given a relation R and two cut vectors* Φ *and* Ψ*, if* $\Phi \gg \Psi$ *then for any subsets* B, C $\subseteq$ A, $R_\Psi|$ = B $\longrightarrow$ C *implies* $R_\Phi|$ = B $\longrightarrow$ C *iff* $R_\Psi|$ = B $\longrightarrow$ C *is compatible with* Φ *in R.*

Proof. ($\Longleftarrow$) Assume that $R_\Psi|$ = B $\longrightarrow$ C is compatible with Φ in R, and $R_\Phi|$ = B $\longrightarrow$ C. Then, for any x, y $\in$ R, if, for every b $\in$ B, $[x(b)]_\Psi = [y(b)]_\Psi$, then, for every c $\in$ C, $[x(c)]_\Psi = [y(c)]_\Psi$. Assume that for every b $\in$ B, $[x(b)]_\Phi = [y(b)]_\Phi$, There are two cases.

Case 1. If for every b $\in$ B, $[x(b)]_\Psi = [y(b)]_\Psi$, then by the assumption that $R_\Psi|$ = B $\longrightarrow$ C, for every c $\in$ C, $[x(c)]_\Psi = [y(c)]_\Psi$, so $[x(c)]_\Phi = [y(c)]_\Phi$.

Case 2. If there is a b $\in$ B such that $[x(b)]_\Psi \neq [y(b)]_\Psi$ then by (*), for every c $\in$ C, $[x(c)]_\Phi = [y(c)]_\Phi = [v_c]_\Phi = ss[v_c]_{\alpha_c}$.

($\Longrightarrow$) Assume that $R_\Psi|$ = B $\longrightarrow$ C is not compatible with Φ in R. Then, there is an x $\in$ R such that for any ($v_c \in D_c$: c $\in$ C), (*) does not hold. Then, as in Example 7, we can construct R, Φ, and Ψ such that $\Phi \gg \Psi$, $R_\Psi|$ = B $\longrightarrow$ C implies $R_\Phi| \neq$ B $\longrightarrow$ C. □

4. Experimental Analysis

4.1. Data Availability. We use the experimental data for heart disease dataset in UCI database, which can be downloaded from the website http://archive.ics.uci.edu/ml/datasets/Heart+Disease. Concept lattice tools ConExp1.3 and lattice miner can be downloaded from the websites: https://sourceforge.net/projects/conexp/ and https://sourceforge.net/projects/lattice-miner/.

4.2. Data Preparation. The heart disease dataset in UCI database contains 303 objects and 14 available attributes. The main purpose of this paper is to analyze the connections between association rules in an original relation and ones in rough relations. To both verify the validity of the method and reduce the computational complexity, we select 24 objects and 5 attributes, as shown in Table 4. The 5 attributes are age: age in years; trestbps: resting blood pressure (in mm Hg on admission to the hospital); restecg: resting electrocardiographic results: Value 0: normal; Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV); Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria; thalach: maximum heart rate achieved; cp: chest pain type: value 1: typical angina; value 2: atypical angina; value 3: nonanginal pain; value 4: asymptomatic.

Table 4 can be converted to a binary relation K = (D, T, I), where D is the set of 24 objects, T= {age, trestbps, restecg, thalach, chest pain type}. The elements of T are abbreviated to *a, d, g, h* and *n*, respectively. "*a*" and "*d*" belong to type 2 in Section 2.3. "*g*" and "*n*" belong to type 1 in Section 2.3. "*h*" belongs to type 3 in Section 2.3, which is necessary to use the FKM clustering.

4.3. Experimental Procedure and Rule Acquisition. We made two groups experiments. The first experiment firstly analyzed hierarchies of attribute domains and then generalized to different levels of attribute values, in order to control the size of the concept lattice. The second experiment used fuzzy attribute values to control the size of the concept lattice.

The First Experiment. Using the concept lattice tool ConExp1.3, which can be downloaded from the website https://sourceforge.net/projects/conexp/, we extract association rules from context. The general form is "<N>P=[C] =><N' >C'", where N is the number of objects satisfying the premise, P is a precondition, C is the confidence of

Table 4: Heart disease set (partly).

No. of patients	age	trestbps	restecg	thalach	chest pain type
1	63	145	2	150	0
2	67	160	2	108	2
3	67	120	2	129	1
4	37	130	0	187	0
5	58	132	2	173	3
6	60	130	2	132	4
7	40	110	2	114	3
8	71	160	0	162	0
9	67	125	0	163	3
10	66	120	2	151	0
11	34	118	2	174	0
12	63	150	2	154	4
13	55	160	2	145	4
14	64	120	2	96	3
15	51	140	0	173	1
16	58	100	2	122	0
17	70	160	0	112	3
18	53	142	2	111	0
19	57	152	0	88	1
20	56	132	2	105	1
21	55	180	1	117	2
22	76	140	1	116	0
23	55	128	1	130	3
24	58	114	1	140	4

Table 5: Association rules from concept lattice in Figure 1.

No.	Rules
1	<3> age2 =[100%]=><3> xueya1
2	< 4 > xueya3 =[75%]=><3> age6
3	<9> age5 =[67%]=><6> xueya1
4	<6> xueya2 =[50%]=><3> age5
5	<6> age6 =[50%]=><3> xueya3
6	<6> age6 =[50%]=><3> xueya1
7	<13> xueya1 =[46%]=><6> age5

Table 6: Association rules from concept lattice in Figure 2.

No.	Rules
1	<3> age2 =[100%]=><3> xueya1
2	<5> age4 =[80%]=><4> xueya2
3	<9> age5 =[67%]=><6> xueya1
4	<6> age6 =[50%]=><3> xueya2
5	<6> age6 =[50%]=><3> xueya1
6	<13> xueya1 =[46%]=><6> age5

association rules, N' is the number of objects meeting the premise, and, in conclusion, and C' is the conclusion. For example, the second association rule in Table 5, i.e., <4> xueya3 =[75%]=><3> age6, meaning that there are 4 objects satisfying the premise for the second level of blood pressure and three objects among them also meet the elderly early old age, and its confidence is 75%.

(1) The age values and blood pressure values are divided into 8 and 4 categories, respectively. Thus, we obtain a relation, called A8BP4. The concept lattice of A8BP4 is shown in Figure 1, from which we obtain some association rules, as shown in Table 5.

(2) The age values and blood pressure values are divided into 8 and 2 categories, respectively. Thus, we obtain a relation, called A8BP2. The concept lattice of A8BP2 is shown in Figure 2. From Figure 2, we obtain some association rules, as shown in Table 6.

(3) The age values and blood pressure values are divided into 3 and 4 categories, respectively. Thus, we obtain a relation, called A3BP4. The concept lattice of A3BP4 is shown in Figure 3. From Figure 3, we obtain some association rules, as shown in Table 7.

(4) The age values and blood pressure values are divided into 3 and 2 categories, respectively. Thus, we obtain a relation, called A3BP2. The concept lattice of A3BP2 is shown in Figure 4. From Figure 4, we obtain some association rules, as shown in Table 8.

For analyzing the concept lattices, we followed the methods of Lei et al. 2016 [16]. Table 9 describes the number of concepts, edges, height, width, and rules in Figures 1–4.

From the description above, we have the following results: (1) the higher the value levels are, the smaller the complexity

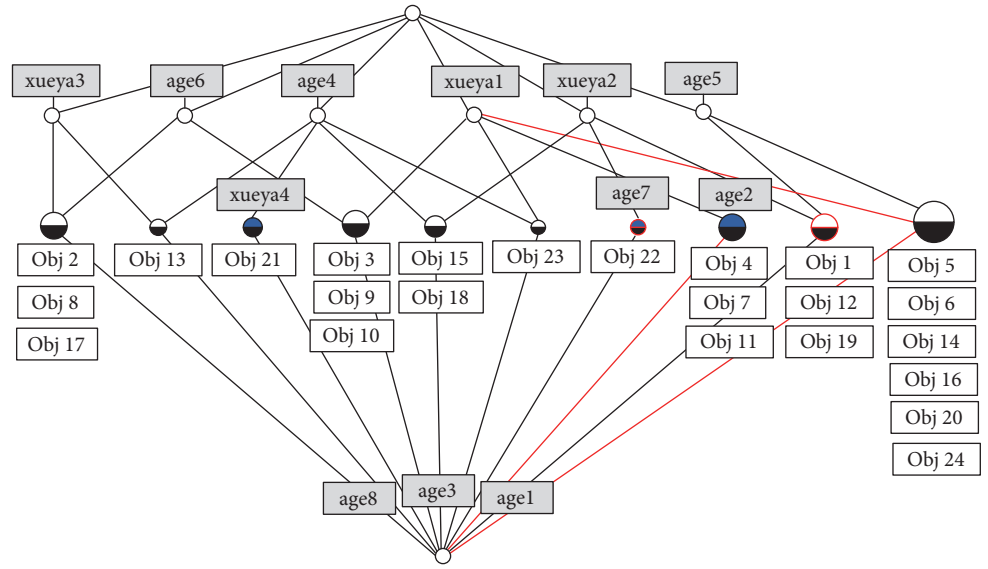


Figure 1: The concept lattice of A8BP4.

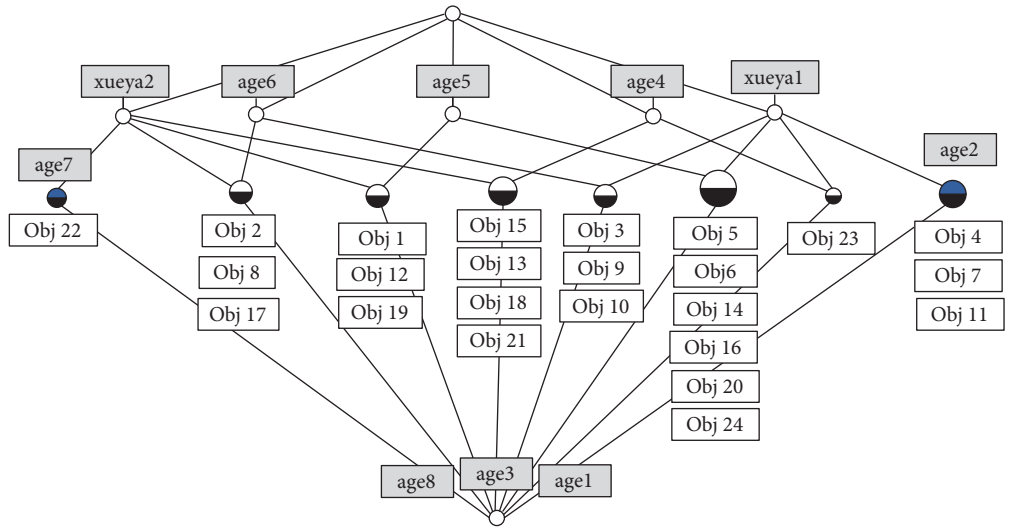


Figure 2: The concept lattice of A8BP2.

Table 7: Association rules from concept lattice in Figure 3.

No.	Rules
1	<3> age1 =[100%]=><3> xueya1
2	<6> xueya2 =[83%]=><5> age2
3	<4> xueya3 =[75%]=><3> age3
4	<13> xueya1 =[54%]=><7> age2

Table 8: Association rules from concept lattice in Figure 4.

No.	Rules
1	<3> age1 =[100%]=><3> xueya1
2	<11> xueya2 =[64%]=><7> age2
3	<7> age3 =[57%]=><4> xueya2
4	<13> xueya1 =[54%]=><7> age2
5	<14> age2 =[50%]=><7> xueya2

of constructing concept lattice is, and the less the association rules are generated; (2) association rules often vary according to the level of abstraction of attribute values. The finer the granularity of value abstraction is, the more the general rules are, and the more the detailed rules are; and (3) the method can reduce relation, control the size of the concept lattice, and satisfy the purpose of different users. In addition, there is a certain relationship among the association rules, as shown in Table 10.

In Table 10, there are some rules which can be obtained at some value level, but not at the other level. There are some rules of implication when the fuzzy processing to different

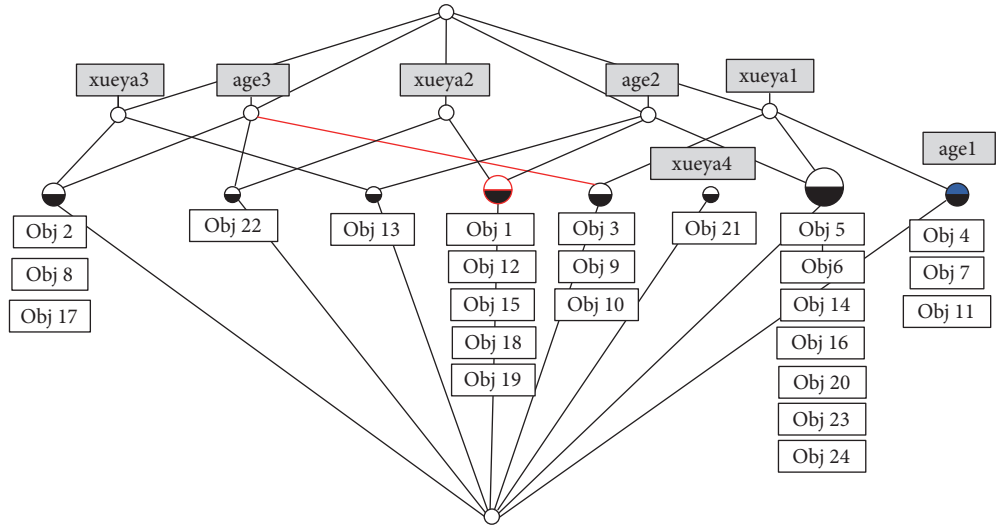


Figure 3: The concept lattice of A3BP4.

Table 9: Comparison of concept lattice from Figures 1 to 4.

	concepts	edges	height	width	rules
Figure 1	18	33	3	[10, 15]	39
Figure 2	15	27	3	[8, 12]	29
Figure 3	15	27	3	[8, 12]	25
Figure 4	11	18	3	[5, 8]	15

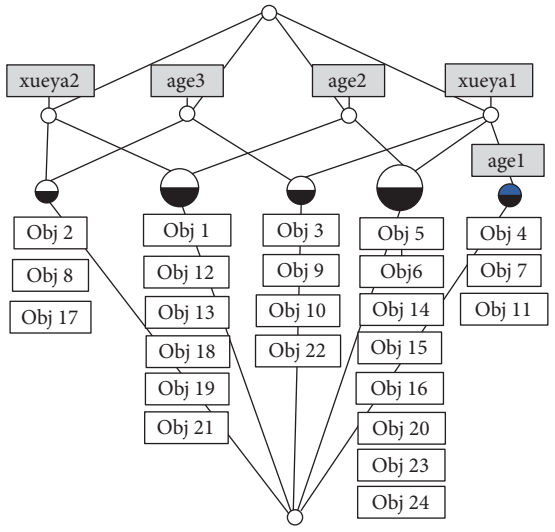


Figure 4: The concept lattice of A3BP2.

levels. For example, rule 3 contains rule 1; rule 4 contains rule 2.

The Second Experiment. This experiment is used to reduce the size of the context and control the number of concepts and improve the quality of the concept. By using the FKM algorithm, we can obtain some fuzzy values of attributes and hence obtain some rough relations from an original relation. Generally, the most common is to classify the age values into three categories and the blood pressure values into two categories. Therefore, Table 4 is converted to a binary relation, as shown in Table 11. The attributes are as follows: a1: young people, a2: middle-aged, a3: elderly, d1: hypertensive, d2: normotensive, g0: normal, g1: having ST-T wave abnormality, g2: probable or definite left ventricular hypertrophy by Estes' criteria, h1: maximum heart rate of low, h2: maximum heart rate of medium, and h3: maximum heart rate of high value, and n0-n4 are heart disease severity, 0 indicates normal, and 1 to 4 indicates serious degree, respectively.

By using lattice miner 1.4, which can be downloaded from the following website: https://sourceforge.net/projects/lattice-miner/, we obtain the concept lattice of Table 11, as shown in Figure 5. It contains 113 concepts and 283 edges, and its height is 6. We define the minimum support and the minimum confidence as 50% and 75%, respectively. Thus, we extract more than 900 rules, as shown in Table 12 (partly).

We select rules with larger values of support and confidence, as shown in Table 13.

To visualize these rules, we use column chart to show the support and confidence of the rules. The support and confidence of the rules in Table 13 are illustrated in Figure 6.

We can get a lot of useful information from Table 13. For example, rule 6 illustrates that the elderly have the larger probability of high blood pressure, and rule 3 illustrates that the elderly maximum heart rate is small. If we put a few rules together, we will obtain more valuable knowledge. For example, we can get a3 □ d1 from rules 6 and 12. From rules

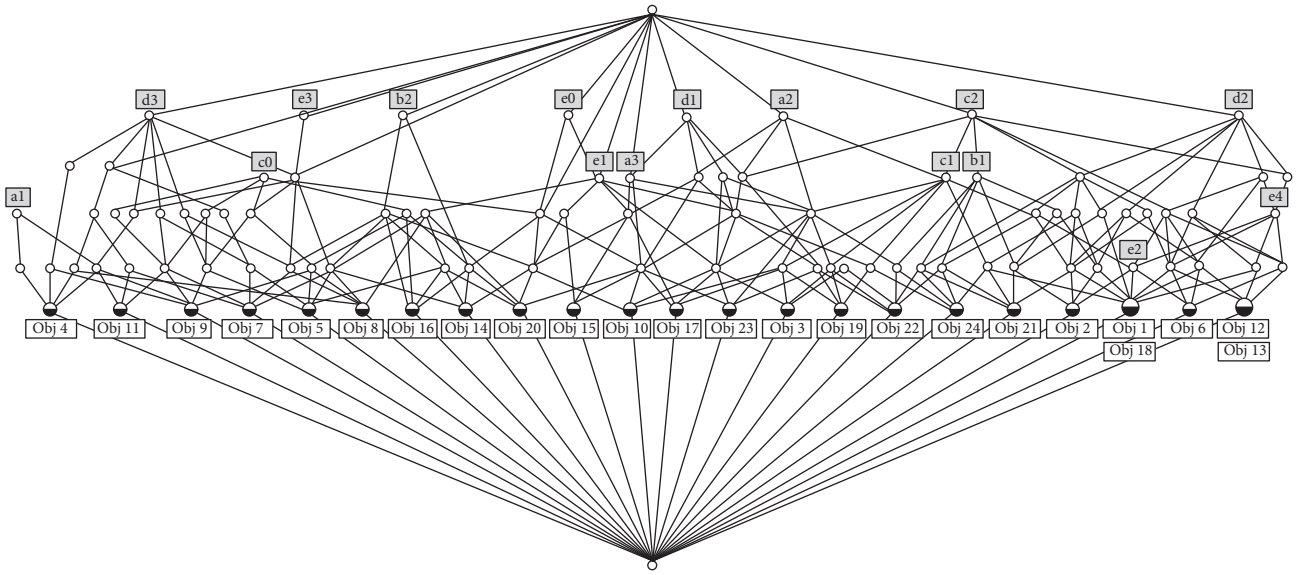


Figure 5: The concept lattice of Table 11.

Table 10: Part of the rules and instructions from Tables 5 to 8.

No.	Rules
1	<4> xueya3 =[75%]=><3> age6
2	<6> age6 =[50%]=><3> xueya2
3	<4> xueya3 =[75%]=><3> age3
4	<7> age3 =[57%]=><4> xueya2

Table 11: The binary relation from Table 4.

	a1	a2	a3	d1	d2	g0	g1	g2	h1	h2	h3	n0	n1	n2	n3	n4
1		+		+				+		+		+				
2			+	+				+	+					+		
3			+		+			+	+				+			
4	+				+	+					+	+				
5		+			+			+			+				+	
6		+			+			+		+						+
7	+				+			+	+						+	
8			+	+		+					+	+				
9			+		+	+					+				+	
10			+		+			+		+		+				
11	+				+			+			+	+				
12		+		+				+		+						+
13		+		+				+		+						+
14		+			+			+	+						+	
15		+		+		+					+		+			
16		+			+			+			+	+				
17			+	+		+				+					+	
18		+		+				+		+		+				
19		+		+		+			+				+			
20		+			+			+			+		+			
21		+		+			+		+					+		
22			+	+			+		+			+				
23		+			+		+		+						+	
24		+			+		+			+						+

Table 12: The part of rules from Figure 5.

No.	Rules	Support	Confidence
R1	a3⟹g1	88.11%	98.88%
R2	a3⟹n4	85.47%	95.92%
R3	a3⟹h1	85.47%	95.92%
R4	a3⟹n2	79.2%	88.88%
R5	a3⟹n3	79.2%	88.88%
R6	a3⟹d1	63.03%	70.74%
R7	h1⟹g1	77.55%	99.57%
R8	a1⟹g1	92.73%	98.59%
R9	a1⟹n4	90.09%	95.78%
R10	a1⟹n2	82.17%	87.36%
R11	a1⟹n3	83.16%	88.42%
R12	d1⟹a3	63.03%	92.71%
R13	d1⟹h1	53.13%	78.15%
R14	a3∧g1⟹n4	84.81%	96.25%
R15	h1∧n2⟹a3	64.68%	92.89%
R16	g1∧n4⟹a1	89.1%	94.07%
R17	h1∧n3⟹g1	71.94%	99.54%
R18	a1∧a3⟹g1	82.17%	98.8%
R19	a3∧n1⟹g1	70.95%	98.62%
R20	a1∧h1⟹g1	72.27%	99.72%
R21	g1∧h1∧n2⟹n4	66.99%	96.66%
R22	a1∧g1⟹n4	89.1%	96.08%
R23	g1∧h1∧n2⟹a3	64.35%	92.85%
R24	g1∧n2∧n3∧n4⟹a3	65.67%	91.28%

Table 13: The part of rules of higher support and confidence.

No.	Rules	Support	Confidence
R1	a3⟹g1	88.11%	98.88%
R2	a3⟹n4	85.47%	95.92%
R3	a3∧n1⟹g1	70.95%	98.62%
R4	a3⟹n3	79.2%	88.88%
R5	g1∧h1∧n2⟹a3	64.35%	92.85%
R6	a1⟹g1	92.73%	98.59%
R7	a1⟹n4	90.09%	95.78%
R8	a1⟹n2	82.17%	87.36%
R9	a1⟹n3	83.16%	88.42%
R10	d1⟹a3	63.03%	92.71%
R11	d1⟹h1	53.13%	78.15%
R12	g1∧n2∧n3∧n4⟹a3	65.67%	91.28%

(2, 4, and 5) and rules (9, 10, and 11), we have the following result: in the current data, the young people and elderly have little difference in heart disease.

5. Conclusions

In order to process natural texts for storing mobile generated data, we firstly extract some formal objects and attributes using NLP, secondly translate the texts into relations, and thirdly process the relations using FCA. In this paper, we mainly discuss the third step. In order to reduce the number of association rules, we propose a method based concept lattice and attribute analysis. This paper establishes the connection between the functional dependencies in an original relation R and corresponding rough relations, proposes the method for extracting reducts in R, and demonstrates the implementation of proposed method on an application in data mining of associative rules. By using rough-values of attributes, we can control the number of concepts and hence improve the quality of the concept. Our experiments show that the method is feasible and effective, which can be applied directly to regular mobile data such as spatial locations and

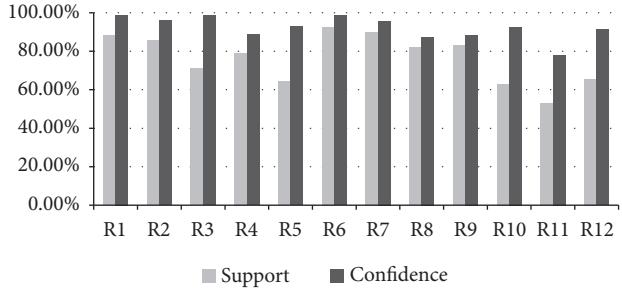


Figure 6: Column chart of support and confidence for rules in Table 13.

healthcare data. In the mobile computing, associative rules can provide potential and useful information for mobile clients.

In the mobile environment, there are the following interesting problems: (1) how to automatically translate natural texts into relations; (2) how to analyze those relations with columns having null values or more complex information; and (3) how to precisely capture mobile users' interests, in order to automatically provide corresponding recommended services.

Data Availability

We use the experimental data for heart disease dataset in UCI database, which can be downloaded from the website http://archive.ics.uci.edu/ml/datasets/Heart+Disease. Concept lattice tools can be downloaded from the following websites: https://sourceforge.net/projects/lattice-miner/ or https://sourceforge.net/projects/conexp/. If readers are interested in our results, they can contact us by email.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is partly supported by the Natural Science Foundation of China under Grants nos. 61572284 and 61502272 and the Promotive Research Fund for Excellent Young and Middle-Aged Scientists of Shandong Province under Grants nos. BS2014DX004 and BS2014DX005.

References

[1] T. Hao, X. Pan, Z. Gu, Y. Qu, and H. Weng, "A pattern learning-based method for temporal expression extraction and normalization from multi-lingual heterogeneous clinical texts," *BMC Medical Informatics and Decision Making*, vol. 18, 2018.

[2] B. Ganter and R. Wille, *Formal Concept Analysis: Mathematical Foundations*, Springer, Berlin, Germany, 1999.

[3] J. Poelmans, D. I. Ignatov, S. O. Kuznetsov, and G. Dedene, "Formal concept analysis in knowledge processing: A survey on applications," *Expert Systems with Applications*, vol. 40, no. 16, pp. 6538–6560, 2013.

[4] P. K. Singh, "Interval-Valued Neutrosophic Graph Representation of Concept Lattice and Its (α, β, γ)-Decomposition," *Arabian Journal for Science and Engineering*, vol. 43, no. 2, pp. 723–740, 2018.

[5] P. K. Singh, C. Aswani Kumar, and J. Li, "Knowledge representation using interval-valued fuzzy formal concept lattice," *Soft Computing*, vol. 20, no. 4, pp. 1485–1502, 2016.

[6] C. A. Kumar, "Fuzzy clustering-based formal concept analysis for association rules mining," *Applied Artificial Intelligence*, vol. 26, no. 3, pp. 274–301, 2012.

[7] K. Raza, "Formal concept analysis for knowledge discovery from biological data," *International Journal of Data Mining and Bioinformatics*, vol. 18, no. 4, pp. 281–300, 2017.

[8] X. Tu, Y. Wang, M. Zhang, and J. Wu, "Using Formal Concept Analysis to Identify Negative Correlations in Gene Expression Data," *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 13, no. 2, pp. 380–391, 2016.

[9] P. K. Singh, A. K. Cherukuri, and J. Li, "Concepts reduction in formal concept analysis with fuzzy setting using Shannon entropy," *International Journal of Machine Learning and Cybernetics*, vol. 8, no. 1, pp. 179–189, 2017.

[10] S. M. Dias and N. J. Vieira, "Concept lattices reduction: Definition, analysis and classification," *Expert Systems with Applications*, vol. 42, no. 20, pp. 7084–7097, 2015.

[11] C. Aswani Kumar, S. M. Dias, and N. J. Vieira, "Knowledge reduction in formal contexts using non-negative matrix factorization," *Mathematics and Computers in Simulation*, vol. 109, pp. 46–63, 2015.

[12] Y. Lei, Y. Sui, and B. Cao, "Formal concept analysis in hybrid relational databases," *International Review on Computers and Software*, vol. 7, no. 6, pp. 2904–2910, 2012.

[13] J. Y. Yu and L. Gan, "FCA application in wireless sensor network," *Journal of Xinyang Agricultural College*, vol. 16, no. 3, pp. 123–126, 2006.

[14] F. Fkih and M. N. Omri, "IRAFCA: an O(n) information retrieval algorithm based on formal concept analysis," *Knowledge and Information Systems*, vol. 48, no. 2, pp. 465–491, 2016.

[15] T. Qian, L. Wei, and J. Qi, "Decomposition methods of formal contexts to construct concept lattices," *International Journal of Machine Learning and Cybernetics*, vol. 8, no. 1, pp. 95–108, 2017.

[16] V. Torra and S. Miyamoto, "A definition for I-fuzzy partitions," *Soft Computing*, vol. 15, no. 2, pp. 363–369, 2011.

[17] Y. Djouadi, B. Alouane, and H. Prade, "Fuzzy clustering for finding fuzzy partitions of many-valued attribute domains in a concept analysis perspective," in *Proceedings of the Joint 2009 International Fuzzy Systems Association World Congress, IFSA 2009 and 2009 European Society of Fuzzy Logic and Technology Conference, EUSFLAT 2009*, pp. 420–425, Portugal, July 2009.

[18] C. Aswani Kumar and S. Srinivas, "Concept lattice reduction using fuzzy K-Means clustering," *Expert Systems with Applications*, vol. 37, no. 3, pp. 2696–2704, 2010.

[19] Y. Lei, J. Tian, and F. Jiang, "Two FCA-based methods for extracting concepts and corresponding concept lattices from hybrid relations," *International Journal of Simulation: Systems, Science and Technology*, vol. 17, no. 27, pp. 35.1-35.12, 2016.

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 9839432, 13 pages
https://doi.org/10.1155/2018/9839432

Research Article

Automatic Approach of Sentiment Lexicon Generation for Mobile Shopping Reviews

Jun Feng[1], **Cheng Gong**[1], **Xiaodong Li**[1] and **Raymond Y. K. Lau**[2]

[1]*College of Computer and Information, Hohai University, Nanjing 211100, China*
[2]*Department of Information Systems, City University of Hong Kong, Hong Kong*

Correspondence should be addressed to Xiaodong Li; xiaodong.c.li@outlook.com

Received 30 March 2018; Revised 11 July 2018; Accepted 30 July 2018; Published 12 August 2018

Academic Editor: Zhe He

The dramatic increase in the use of smartphones has allowed people to comment on various products at any time. The analysis of the sentiment of users' product reviews largely depends on the quality of sentiment lexicons. Thus, the generation of high-quality sentiment lexicons is a critical topic. In this paper, we propose an automatic approach for constructing a domain-specific sentiment lexicon by considering the relationship between sentiment words and product features in mobile shopping reviews. The approach first selects sentiment words and product features from original reviews and mines the relationship between them using an improved pointwise mutual information algorithm. Second, sentiment words that are related to mobile shopping are clustered into categories to form sentiment dimensions. At each sentiment dimension, each sentiment word can take the value of 0 or 1, where 1 indicates that the word belongs to a particular category whereas 0 indicates that it does not belong to that category. The generated lexicon is evaluated by constructing a sentiment classification task using several product reviews written in both Chinese and English. Two popular non-domain-specific sentiment lexicons as well as state-of-the-art machine-learning and deep-learning models are chosen as benchmarks, and the experimental results show that our sentiment lexicons outperform the benchmarks with statistically significant differences, thus proving the effectiveness of the proposed approach.

1. Introduction

With the rapid development of smartphones, mobile shopping, which is already popular, is expected to grow faster. After shopping, people provide a large number of reviews about different kinds of products on the Internet. Different products could be preferred by different consumer groups. Hence, it is becoming increasingly important to learn about a customer's emotional inclinations and favorite products through online reviews. Sentiment classification can be performed using machine-learning, lexicon-based, and hybrid approaches. Sentiment lexicons are important resources for these approaches. The analysis of sentiment orientation is widely known as a domain-specific task. However, almost all the existing sentiment lexicons are general lexicons, which are not suitable for the analysis of product reviews on the Internet. Thus, automatic construction methods for sentiment lexicons have attracted increasing attention recently, especially methods for constructing sentiment lexicons aimed at mobile shopping.

Sentiment analysis, which is also called opinion mining, review mining (appraisal extraction), or attitude analysis, is the task of detecting, extracting, and classifying opinions, sentiments, and attitudes concerning different topics [1]. In a machine-learning approach, sentiment analysis can be considered as a supervised classification task. Pang et al. [2] solved the sentiment classification problem by training the classifier. However, most machine-learning approaches rely on features that are engineered by machine-learning methods. In a lexicon-based approach, a dictionary is created to judge whether the polarity of words in the text is positive or negative. For example, Turney [3] scanned a review for phrases that matched certain patterns (adjectives and adverbs) and then added up all sentiment orientations to compute the orientation of a document. A hybrid approach combines both the above approaches and has a relative

advantage in sentiment analysis. Ortigosa et al. [4] developed a lexicon from a corpus and then chose sentiment words along with the labeled class as the input features for a machine-learning classification method. Sentiment lexicons play a key role in a majority of the above methods.

A sentiment lexicon (or an opinion lexicon) is a list of words and phrases that are commonly used to express positive or negative sentiments [5]. Researchers have proposed many approaches to compile these sentiment words. Technically, the existing automatic lexicon construction methods for both English and Chinese languages are mainly divided into corpus-based and knowledge-based methods. Turney [3] developed a corpus-based method in which the sentiment orientation of a word was judged by using pointwise mutual information (PMI) to describe the closeness of the word and seed words. Knowledge-based methods require a relatively complete knowledge base. Hu and Liu [6] constructed a sentiment lexicon by searching for the synonyms and antonyms of a word in WordNet. For a specific domain, the sentiment lexicon constructed from the corresponding domain corpus is more practical. When building a sentiment lexicon for online product reviews, the product features modified by sentiment words are also very important factors [7]. However, the existing general sentiment lexicons usually include only limited common words, and these words are divided into binary or other fixed categories according to the sentiment orientation.

In this paper, we present a novel method to construct a domain-specific sentiment lexicon by mining the relationship between sentiment words and product features in a specific corpus. In our approach, first, a sentiment matrix is constructed based on the relationship between sentiment words and product features. Every row of the sentiment matrix is regarded as a vector representation of the sentiment word. The sentiment words in the matrix space are clustered based on the distance between the vectors. Second, sentiment words that are related to mobile shopping are clustered into categories to form sentiment dimensions. In the process of building the sentiment matrix, the idea of term frequency–inverse document frequency (TFIDF) is utilized to screen the product features. Furthermore, the traditional PMI algorithm is improved to obtain a new algorithm called EPMI, which is more suited to mobile shopping reviews. Extensive experiments are performed on seven different domain product reviews, which include reviews in both Chinese and English. Compared to two popular general lexicons as well as state-of-the-art machine-learning and deep-learning models, our lexicon can obtain satisfactory classification performance. The experimental results also show that the filtering of product features and the application of the EPMI algorithm can greatly improve the performance of our lexicon for mobile shopping reviews.

The rest of the paper is structured as follows. Discussions on sentiment classification and lexicon generation and a review of the most recent research are presented in Section 2. Our methods for constructing the sentiment lexicon for mobile shopping reviews and a walk-through example of our methods are presented in Section 3. The experimental setup and results are described in Section 4. The conclusions of the paper are summarized in Section 5.

2. Related Work

This section is structured as follows. In the first part of this section, we review previous works on sentiment classification approaches. In the second part, we summarize works on approaches for sentiment lexicon creation. In addition, we briefly introduce the sentiment dimensions considered in the lexicon and product feature identification for product reviews.

2.1. Sentiment Classification. Sentiment classification aims to automatically classify the text of reviews written by customers into positive or negative opinions. Sentiment classification techniques can be roughly divided into machine-learning, lexicon-based, and hybrid approaches [8].

Machine-Learning Approaches. In such approaches, the analysis of customers' emotional inclinations is considered to be a problem of polarity classification. Pang et al. [2] applied three machine-learning methods (naive Bayes (NB), maximum entropy, and a support vector machine (SVM)) to sentiment classification as a form of traditional topic-based categorization. Zhang et al. [9] used machine learning (NB and SVM) to classify the sentiments expressed in restaurant reviews written in Cantonese. Li et al. [10] adopted extreme learning machine and deep-learning architecture to improve feature representations for text classification. Enríquez et al. [11] showed how a vector-based word representation obtained via Word2Vec can help in improving the results of a document classifier based on the bag-of-words model. However, these supervised machine-learning techniques require a large corpus of training data, and their performance is acceptable only if the match between the training and test data is good.

Lexicon-Based Approaches. These approaches adopt a lexicon to perform sentiment analysis by counting and weighting sentiment words that have been evaluated and tagged [12]. Nasukawa and Yi [13] developed a method to determine subject favorability by creating a sentiment lexicon containing 3513 sentiment terms. Qiu et al. [14] used a lexicon-based approach to identify sentiment sentences in contextual advertising. The most common lexicon resources are SentiWordNet, WordNet, and ConceptNet, and among these resources, SentiWordNet is the most widely used [15].

Hybrid Approaches. Nowadays, researchers are also using combined approaches, in which two or more approaches are combined to achieve better accuracy. Sindhwani and Melville [16] presented a unified framework in which lexical background information, unlabeled data, and labeled training examples can be effectively combined. Li et al. [17] set up a system to analyze the market impact by combining the stock price and news sentiment. Ortigosa et al. [4] performed sentiment classification and sentiment change detection on Facebook comments using a hybrid approach. They combined lexicon-based and machine-learning methods by considering a lexicon as the source of features and using a classification model to evaluate the lexicon; this approach is similar to the one used in our experiments in this study.

2.2. Lexicon Creation. A sentiment lexicon is an important tool for identifying the sentiment polarity of reviews provided by mobile users [18]. Two methods are commonly used to generate sentiment lexicons: knowledge-based and corpus-based methods.

Knowledge-Based Methods. These methods exploit available lexicographical resources such as WordNet or HowNet. Hu and Liu [6] developed a lexicon by searching for the synonyms and antonyms of a word in WordNet. Kamps [19] inferred that the greater the closeness of two words, the smaller the number of iterations required to determine the synonymous relationship between the words. Both these studies used the relationship between words in a knowledge base. The main strategy in these methods is to first manually collect an initial seed set of sentiment words and their orientations and then search for their synonyms and antonyms in a knowledge base to expand this set [12]. However, very few complete and robust knowledge bases are available for the Chinese language.

Corpus-Based Methods. These methods depend on syntactic patterns or patterns that occur together along with a seed list of opinion words to find other opinion words in a large corpus [20]. Hatzivassiloglou and McKeown [21] found that, with a change in the emotional polarity in the text, the turning point appears but concatenation does not. Based on the idea that the emotional polarity of a word tends to be consistent with the emotional polarity of its neighboring words, Turney and Littman [22] constructed a dictionary from a large corpus. Both these works [21, 22] are based on a corpus rather than a knowledge base. The corpus-based approach has a major advantage in that it can find domain-specific words and their orientations if a domain-specific corpus is used in the discovery process. Therefore, our work also focuses on a corpus-based approach. In addition, PMI is commonly used in this approach to exploit the syntactic patterns of cooccurrence patterns. Turney and Littman [22] used PMI and latent semantic analysis to measure the correlation between two words, and this method, which uses PMI to calculate the correlation between a word and seed word, is called semantic-orientation PMI (SO-PMI). Yang et al. [23] introduced a method based on SO-PMI to construct a sentiment lexicon and improved the SO-PMI model based on user behavior. In the process of our lexicon construction, we improve the traditional PMI to make it more suitable for mobile shopping reviews.

In the process of lexicon construction, we focus on two issues: the sentiment dimensions of the lexicon, and feature or topic identification in product review domains.

Sentiment Dimensions. Ekman [24] found that humans have six basic emotional categories: happiness, sadness, fear, surprise, anger, and jealousy. Ekman's theory, which is accepted by numerous psychologists and linguists, is widely used in the field of sentiment analysis. Rubin et al. [25] presented an empirically verified model on the basis of the idea [26] that an emotion can be divided into eight categories with two major bipolar dimensions: positive and negative effects.

Table 1: Mathematical symbols and their meanings.

Symbol	Meaning
A, B, F	Sentiment matrix
C	Matrix of relationship between features
$E = \{e_1, e_2, \ldots, e_n\}$	Set of sentiment words
$M = \{m_1, m_2, \ldots, m_p\}$	Set of primitive product features
$M' = \{m'_1, m'_2, \ldots, m'_t\}$	Set of product features after filtering
$D = \{d_1, d_2, \ldots, d_n\}$	Set of reviews

Although early approaches simply focused on this binary classification [27], we not only consider the two polarities but also anticipate that sentiment words can be reasonably clustered into finer-gained categorizations.

Feature Identification. Considering that many words in different fields may have different sentiment polarities, it is necessary to explicitly extract the sentiment words and topics or product features, especially in the mobile review domain. Fast et al. [28] found out that using experts or crowdsourcing to construct domain-specific sentiment lexicons is very difficult. Zhang et al. [29] proposed a hybrid method that combined Apriori and PMI to extract product features. Mishne [30] chose the part of speech (POS) and word counts as features in a text classification task. In our research, the primitive product feature extraction also uses the POS as a selection criterion.

3. Methods

In this section, we present our proposed framework to generate domain-specific sentiment lexicons for mobile shopping. Figure 1 shows the framework of our method. The domain-specific lexicon is based on the relationship between sentiment words and product features modified by the sentiment words. A sentiment matrix is adopted to represent the relationship between the sentiment words and product features. First, we use PMI to express the relationship between sentiment words and product features. Second, we use TFIDF to filter product features so as to reduce the matrix dimension. Finally, we improve the traditional PMI to develop a new algorithm called EPMI, which is used to build a new sentiment matrix. Each row in the sentiment matrix is a vector representation of the sentiment word. After obtaining the sentiment matrix, we cluster the sentiment words into several categories based on the distance between their vector representations. The mathematical symbols used in the process of construction are listed in Table 1.

3.1. Building of Primitive Sentiment Matrix. To perform the key step of mining the relationship between sentiment words and product features, we need to determine the sentiment words and product features in the corpus. To choose the terms from the corpus as candidate words, we use the POS.

Sentiment words are commonly used to express positive or negative sentiments. Sentiment lexicons usually contain such words, which can indicate the sentiment polarity (e.g.,

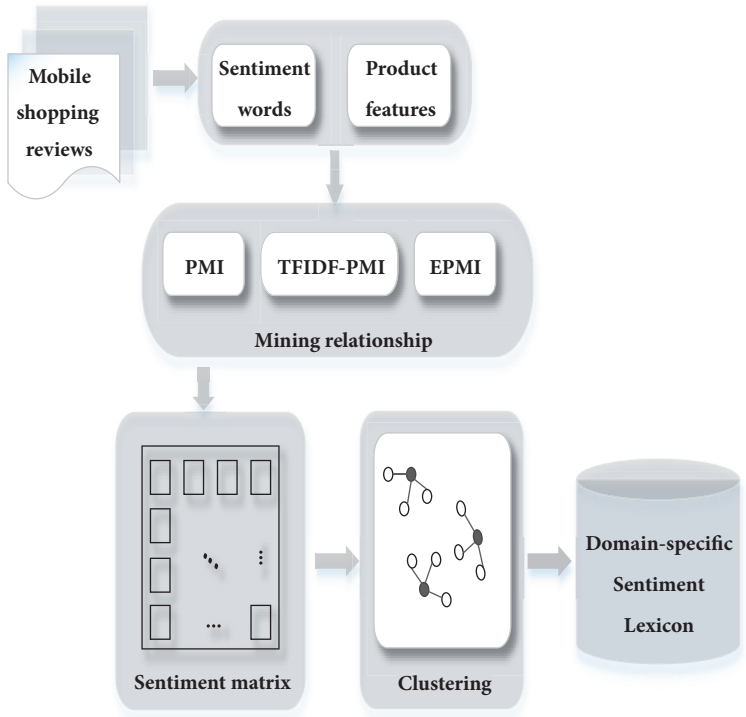


Figure 1: Framework of method used for constructing domain-specific lexicon for mobile shopping.

"good" and "wonderful" indicate positive opinions, whereas "rubbish," "cheap," and "terrible" indicate negative opinions). In mobile shopping reviews, a number of verbs can also indicate the sentiment polarity (e.g., "like" and "love" indicate positive opinions, whereas "dislike" and "refund" indicate negative opinions). In some previous studies [31, 32], the words whose POS is an adjective or adverb are considered as sentiment words. The sentiment lexicons developed or used in some other studies [6, 33] are also mainly concerned with adjectives and adverbs. In addition, product features in the product review domain are usually nouns or noun phrases found in review sentences [6]. Therefore, we choose adjectives, adverbs, and verbs as sentiment words and choose nouns as primitive product features. For instance, in the hotel review "The food in the dining room is really good, the breakfast tastes good," the product features are "dinning," "breakfast," and "food," and the sentiment words are "good" and "tastes."

If a sentiment word A modifies a product feature B, we consider that there is a relationship between them. In mobile shopping reviews, this relationship can be shown as a phenomenon of cooccurrence. We use PMI to quantify this type of cooccurrence relationship. PMI is defined as

$$\mathrm{PMI}\left(word_1, word_2\right) = \log_2\left(\frac{\mathrm{p}\left(word_1, word_2\right)}{\mathrm{p}\left(word_1\right)\mathrm{p}\left(word_2\right)}\right) \quad (1)$$

Here, $\mathrm{p}(word_1, word_2)$ is the cooccurrence probability of $word_1$ and $word_2$ in the local window and is expressed as

$$\mathrm{p}\left(\mathrm{word}_1, \mathrm{word}_2\right) = \frac{\mathrm{count}\left(word_1, word_2\right)}{N} \quad (2)$$

where N is the total number of words contained in the corpus. $\mathrm{count}(word_1, word_2)$ represents the number of occurrences of the two words in the local window. Similarly, the frequency of each word can be obtained as

$$\mathrm{p}\left(\mathrm{word}\right) = \frac{\mathrm{count}\left(word\right)}{N} \quad (3)$$

In (1), $\mathrm{p}(word_1)\mathrm{p}(word_2)$ gives the probability of cooccurrence if these two words are statistically independent. The ratio of $\mathrm{p}(word_1, word_2)$ to $\mathrm{p}(word_1)\mathrm{p}(word_2)$ is thus a measure of the degree of statistical dependence between the words.

The PMI value between the sentiment words and product features can reflect the relationship between them. By calculating the PMI value between all the sentiment words and product features, we can obtain a sentiment matrix that contains the relationship between the sentiment words and product features. Let us denote $E = \{e_1, e_2 \ldots e_n\}$ as the set of sentiment words and $M = \{m_1, m_2, \ldots m_p\}$ as the set of product features. Matrix A, as shown below, consists of n rows and p columns.

Definition 1. Sentiment matrix A: the rows represent the sentiment words, whereas the columns represent the product features. The value of each cell w_{ij} is given by PMI(e_i, m_j).

$$A = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1p} \\ w_{21} & w_{22} & \cdots & w_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{np} \end{bmatrix} \quad (4)$$

In the above matrix, each sentiment word e_i can be represented as a vector $\overrightarrow{e_i} = [w_{i1}, w_{i2}, \ldots, w_{ip}]$. Sentiment matrix A is the primitive sentiment matrix, and this matrix is optimized, as described in the next subsection.

3.2. Filtering of Product Features. So far, we have obtained the primitive sentiment matrix A, and each sentiment word e in the matrix can be represented as a vector. According to our approach, these vectors should be clustered into several categories. However, we found that the number of product features is very large because we consider all nouns as product features. Consequently, the word vector will face the dimension disaster problem. The clustering of high-dimensional data is still a challenging problem because of the curse of dimensionality [34]. In addition, the use of high dimensions will result in low computational efficiency, especially in mobile computing. In Hu and Liu's study [6], only those product features regarding which many people have expressed their opinions are reserved. Similarly, we also select key product features from the primitive nouns. Next, we will describe our feature selection method in detail.

The high-dimension problem stems from the large number of nouns in the corpus. The number is large because we choose all nouns as product features. For instance, consider the product review "This hotel is great, I can recommend my mom to live next time." The word "mom" and "time" will be treated as product features, but these words do not represent any features of the hotel. In addition, this type of nouns can be found everywhere in mobile shopping reviews. Therefore, it is necessary to filter out the key product features rather than choose all nouns as the product features. Product features should be nouns that frequently appear in a particular category of product reviews and rarely appear in other categories. Therefore, we use the idea of TFIDF to select real product features. TFIDF is defined as

$$\text{TFIDF}(word) = \text{TF}(word) \times \text{IDF}(word) \quad (5)$$

Here, TF(*word*) means the term frequency of the word in the document. IDF(*word*) means the inverse document frequency, that is, whether the word is common or rare across all documents. It is important to note that the TFIDF value of the same word may be different in different documents. However, TFIDF is usually used for documents rather than pieces of reviews. There may be thousands of comments about a single product. We just need to merge the same kind of comments together to form the corresponding document.

From (5), we can obtain the TFIDF value of words in different documents. Unlike the analysis described in the previous subsection, here, we choose the nouns whose TFIDF values are relatively high in the document as product features of the product. Unexpectedly, we find that the nouns whose TFIDF values are relatively high happen to be words that are closely related to the reviewed product. For example, if there are numerous reviews about a hotel, we can retrieve words such as "bathroom" and "air-conditioning" from the corresponding document. When we are commenting on a hotel, we often refer to the "bathroom" or "air-conditioning" in the hotel. However, these two words rarely appear in the reviews of products from other domains such as the electronics domain. We can certainly define a threshold α that the TFIDF value of real product features must reach. Let us denote $\mathrm{M}' = \{m'_1, m'_2, \ldots, m'_t\}$ where $\mathrm{M}' \subseteq M$ as the set of the remaining product features after filtering by TFIDF. Accordingly, we can obtain another sentiment matrix that is similar to sentiment matrix A. This matrix (sentiment matrix B) consists of n rows and t columns.

Definition 2. Sentiment matrix B: this matrix can be considered to be part of sentiment matrix A. The rows represent the sentiment words, whereas the columns represent the product features after filtering by TFIDF. The value of each cell w_{ij}, which is the same as that in sentiment matrix A, is given by PMI(e_i, m'_j).

In matrix B, each sentiment word e_i can be represented as a vector $\overrightarrow{e_i} = [w_{i1}, w_{i2}, \ldots, w_{it}]$. Here, t can be considerably less than p when the threshold α is set appropriately. Compared to sentiment matrix A, sentiment matrix B can effectively solve the high-dimension problem in word embedding. However, there are still some defects in the sentiment matrix, which will be elaborated in the next subsection.

3.3. Optimization of Sentiment Matrix by EPMI. Here, we introduce an example from hotel reviews to further explain the defect in sentiment matrices A and B. We focus on two sentiment words (e_1 = "rich" and e_2 = "hearty") and two product features (m_1 = "food" and m_2 = "breakfast"). Both these sentiment words can be used to express opinions about a wide variety of foods. The meanings of these two sentiment words are very similar, and these words are commonly used in the hotel review domain. If we just consider the two features m_1 and m_2, e_1 and e_2 can be represented as $\overrightarrow{e_1} = [w_{11}, w_{12}]$ and $\overrightarrow{e_2} = [w_{21}, w_{22}]$ in the sentiment matrix. w_{ij} is given by PMI(e_i, m_j).

As is well known, the distance or angle between word vectors can be considered to be the similarities between words. The greater the similarity between two words, the shorter the distance between them. However, in a hotel review, the two sentiment words (e_1, e_2) and two product features (m_1, m_2) can be matched with each other flexibly. Although some customers may usually modify m_1 with e_1 and m_2 with e_2, they may rarely modify m_1 with e_2 and m_2 with e_1. This means that the PMI value of (e_1, m_1) and (e_2, m_2) is relatively high, but the PMI value of (e_1, m_2) and (e_2, m_1) is very low. Therefore, an illusion is created that e_1 and e_2 are irrelevant in the two dimensions of m_1 and m_2. This irrational result stems from the flexibility of product reviews and the

Input:
$M = \{m_1, m_2, \ldots, m_p\}, M' = \{m'_1, m'_2 \ldots, m'_t\}, D = \{d_1, d_2, \ldots, d_n\}$
Output: Matrix C[t][p]
(1) i = 1, C[t][p] = zero matrix
(2) **while** $i \leq t$ **do**
(3) j = 1
(4) **while** $j \leq p$ **do**
(5) **for** each $d \in D$ **do**
(6) **if** (m'_i, m_j) *in* d && $m'_i \neq m_j$ **then**
(7) C[i][j] = C[i][j] + 1
(8) j = j + 1
(9) i = i + 1
(10) **For** each row $row \in C[t][p]$ **do**
(11) row = normal(row)
(12) **Return** C

Algorithm 1: Mining relationship between product features.

diversity of vocabulary in mobile shopping reviews. Although e_1 rarely modifies m_2, it cannot be simply considered to be irrelevant.

When we consider the relationship between a sentiment word and product feature, it is not sufficient to just calculate the PMI value of these two words directly. We still need to consider the relationship between the sentiment word and other product features that are related to the initial product feature. In the mobile shopping reviews about a hotel, there are many features related to m_2 (*breakfast*), such as "food" and "dinning." Therefore, when we calculate the PMI value of e_1 and m_2, we consider the cooccurrence of not only e_1 and m_2 but also e_1 and m_1 or other product features related to m_2. We use $u_{ij} \in [0, 1]$ to reflect the degree of correlation between the two product features m_i and m_j. The larger the value of u_{ij}, the more relevant m_j to m_i. $u_{ij} = 0$ indicates that the two features m_i and m_j are irrelevant. In particular, if the features m_i and m_j represent the same feature, the value of u_{ij} between them is zero. Considering all the p product features contained in the corpus, we define EPMI as

$$\text{EPMI}\left(e_i, m_j\right) = \text{PMI}\left(e_i, m_j\right) + \sum_{k=1}^{p} u_{jk} * \text{PMI}\left(e_i, m_k\right) \quad (6)$$

Once we know the value of u_{ij} between any two features m_i and m_j, we can obtain the EPMI value on the basis of the PMI value. Considering that we can screen the features according to the method described in the previous subsection, we focus on the correlation between the remaining product features after filtering and all the primitive features. We assume that the more frequently two features appear in the same review, the higher the correlation between them is. The pseudocode for mining the relationship between them is presented in Algorithm 1.

Following the earlier definitions, M is still the set of all product features for a given kind of production, and M' is the set of product features obtained by carrying out filtering using the approach described in the previous subsection. D is the set of reviews related to the product. (m'_i, m_j) *in* d means that features m'_i and m_j appear together in review d. The function normal() is a simple normalization function that is used to ensure that every element in the vector belongs to $[0, 1]$. This algorithm is an effective algorithm in the sense that it can find the features that are the most relevant to a specific feature. We can obtain matrix C using the above algorithm. After obtaining matrix C, we can use EPMI to build a new sentiment matrix.

Definition 3. Sentiment matrix F can be determined by (7). The only difference between sentiment matrices F and B is that matrix F is obtained using our approach (EPMI) rather than the traditional PMI. In other words, each cell of sentiment matrix F represents the EPMI value between the sentiment words and product features rather than the PMI value between them.

$$F[n][t] = B[n][t] + A[n][p] * C^T[t][p] \quad (7)$$

So far, we have obtained three sentiment matrices A, B, and F. The sentiment words can be represented by the vectors in each of the sentiment matrices.

In mobile shopping reviews, customers often use different sentiment words to modify different product features. In addition, customers may have a good feeling about some features of a product, but they may be dissatisfied with some other features at the same time. Therefore, different product features also reflect different feelings. We assume that sentiment words can be divided into different categories according to the relationship between them and the product features.

Therefore, we cluster the sentiment words into several categories rather than into binary or other fixed categories. In other words, the sentiment dimension of a word in our domain-specific lexicon is flexible rather than having only limited emotional polarity. For each sentiment dimension, each sentiment word can take the value of 0 or 1, where 1 indicates that it belongs to a particular category whereas 0 indicates that it does not belong to that category. If we cluster the sentiment words into five categories, the representation

Table 2: Sample product reviews.

1	早餐 很 丰盛, 餐厅 非常 大。 (Breakfast is very hearty, the dining room is very large)
2	食物 种类 丰富, 那里 有 免费 的 早餐。 (The food is plentiful and there is a free breakfast)
3	餐厅 免费 供应 丰盛 的 早餐。 (The dinning serves a hearty breakfast for free)
4	餐厅 里 的 食物 挺 丰富 (The food in the dinning is rich)
5	餐厅 食物 真不错, 早餐 尝起来 很 好吃 (The food in the dinning is really good, the breakfast tastes good)

Table 3: Number of instances of cooccurrence of features.

	m_1	m_2	m_3	m_4
m_1'	0	2	2	1
m_2'	2	0	3	1

$\overrightarrow{e} = [0, 0, 1, 0, 0]$ means that the word e belongs to the third category. The flexibility of sentiment dimensions is a main characteristic of the domain-specific lexicon built using our approach.

3.4. Walk-Through Example. Here, we will elaborate on the differences between EPMI and PMI using an example. Suppose that we want to determine the semantic correlations between the sentiment words e_1 = "丰富" (rich) and e_2 = "丰盛" (hearty), and the five sentences listed in Table 2 are our corpus. This small corpus is a part of Chinese mobile shopping reviews about hotels.

Here, N is 33 because these five comments contain 33 Chinese words. In this example, there are four nouns (or primitive product features), i.e., $M = \{m_1 \sim m_4 \mid$ "食物" (food), "早餐" (breakfast), "餐厅" (dining room), "种类" (variety)$\}$. To keep our example simple and understandable, we focus on only two features, m_1 = "食物" (food) and m_2 = "早餐" (breakfast). Therefore, M' can be $\{m_1', m_2'\}$. count(e_1) is 2 because the word e_1 appears only twice in this small corpus. Similarly, count(e_2) is 2, count(m_1) is 3, and count(m_2) is 4. The size of the local window of cooccurrence is set as 3. count(e_1, m_1) is 2 because e_1 and m_1 cooccur twice in the second and fourth comments within the window. Similarly, count(e_2, m_2) is 2, count(e_1, m_2) is 0, and count(e_2, m_1) is 0. Using (1), we can obtain PMI(e_1, m_1) = $\log_2(p(e_1, m_1)/p(e_1)p(m_1))$ = $(2/33)/((2/33)(3/33))$ = 3.46 and PMI(e_1, m_2) = 0. Similarly, PMI(e_2, m_1) = 0 and PMI(e_2, m_2) = $(2/33)/((2/33)(4/33))$ = 3.04. Therefore, the two sentiment words e_1 and e_2 can be represented as $\overrightarrow{e_1}$ = [3.46, 0] and $\overrightarrow{e_2}$ = [0, 3.04], respectively, in sentiment matrices A and B.

We calculate u_{ij} using Algorithm 1. First, by iterating through these five comments, we obtain the number of instances of cooccurrence of m_i' and m_j (Table 3). In this table, each cell shows the number of times that two features appear together in the same comment. The values in this table are similar to those in matrix C obtained in Algorithm 1.

Next, we normalize this table or matrix. We choose the min–max normalization function as the normal() function in the algorithm. Finally, we obtain $u_{12} = 1$, $u_{13} = 1$, $u_{14} = 1/2$, $u_{21} = 2/3$, $u_{23} = 1$, and $u_{24} = 1/3$. Note that u_{11} and u_{22} are 0.

Using (6), we obtain EPMI(e_1, m_2) = PMI(e_1, m_2) + u_{21}PMI(e_1, m_1) + u_{22}PMI(e_1, m_2) + u_{23}PMI(e_1, m_3) + u_{24}PMI(e_1, m_4) = 3.65 and EPMI(e_1, m_1) = 5.48. Similarly, EPMI(e_2, m_1) = 6.08 and EPMI(e_2, m_2) = 6.08. The two sentiment words e_1 and e_2 can be represented as $\overrightarrow{e_1}' = [5.48, 3.65]$ and $\overrightarrow{e_2}' = [6.08, 6.08]$ in sentiment matrix F.

It is obvious that the distance between $\overrightarrow{e_1'}$ and $\overrightarrow{e_2'}$ is considerably closer than that between $\overrightarrow{e_1}$ and $\overrightarrow{e_2}$, irrespective of the Euclidean or cosine distance. This result reflects the difference between our EPMI algorithm and PMI. These two sentiment words are very similar and are commonly used in hotel reviews. The greater the similarity between the two words, the shorter the distance between them. In the clustering model, the vectors that are located at shorter distances are easier to be clustered into the same category.

4. Experiments

To evaluate the domain-specific lexicon developed using our approach, we design an experimental setup using which we compare the proposed domain-specific lexicon with two popular general lexicons and with state-of-the-art machine-learning and deep-learning approaches that do not use a lexicon. We mainly evaluate different lexicons and approaches using document-level classification tasks in the domain of online product reviews. For hybrid sentiment classification methods, we consider the features of the document vector representation as the lexicon. We use the F1-measure as our main evaluation index and choose NB and SVM as the classifiers. In the following subsections, the details of the experiments and their results are described.

4.1. Dataset. The dataset includes both Chinese and English shopping reviews. These reviews are for seven types of products. The detailed statistics of this dataset are listed in Table 4.

The Chinese product reviews include three domains: hotel, cloth, and fruit. The hotel reviews are provided by Dr. Tan (http://download.csdn.net/download/lssc4205/9903298), and the cloth and fruit reviews are crawled from a mobile shopping application JD (https://www.jd.com/). The English reviews are obtained from the famous Amazon product review dataset collected by Blizter et al. [35]. It is widely used as a benchmark dataset for cross-domain sentiment classification. Four domains—book, DVD, electronics, and kitchen—are included in this dataset. For each domain, 1000 positive and 1000 negative reviews are included.

4.2. Experimental Design. We use the open-source software jieba (https://pypi.python.org/pypi/jieba/) to carry out preprocessing tasks on the Chinese product reviews, including Chinese word segmentation and POS tagging. For the

Table 4: Statistics of dataset.

Polarity	Chinese			English			
	Hotel	Cloth	Fruit	Books	Kitchen	DVD	Electronics
Positive	5321	5000	5000	1000	1000	1000	1000
Negative	2444	5000	5000	1000	1000	1000	1000

sentiment classification approaches that do not use a lexicon, we compare our method with the classical bag-of-words and deep-learning model Word2Vec [36]. Furthermore, we compare the domain-specific sentiment lexicon with two popular general sentiment lexicons. We use the scikit-learn [37] python library implementation of the classifier. The detailed differences between the three test groups are described below.

(a) No Lexicon

(BOW) The classical method to express the document involves the use of the bag-of-words model [2]. Each document d is represented by a feature-presence vector $\overrightarrow{d} = [0, 1, \ldots, 0]$.

(W2V) In addition to the bag-of-words classical representation, we use the encoding of words provided by Word2Vec, which is a deep-learning tool released by Google in 2013. This tool adopts two main model architectures—the continuous bag-of-words model and continuous skip-gram model—to learn the vector representations of words [38]. To use Word2Vec for document-level tasks, a method is required that can unify all word vectors and generate a single vector representing the entire document [11]. Thus, the final representation is obtained according to the number of words contained in the document as follows:

$$\overrightarrow{v_d} = \frac{\sum_{i=0}^{n} \overrightarrow{v_i}}{n} \tag{8}$$

We use the genism (https://radimrehurek.com/gensim/models/word2vec.html) python library implementation of Word2Vec. We use the default values for almost all the parameters and use vectors with 200 dimensions.

(b) General Lexicon

For this test group, we use the hybrid sentiment classification approaches. That is, we consider the words in the lexicon, the sentiment dimensions of the lexicon, and a combination of the words and dimensions as the features for the machine-learning classifier. First, we choose a general sentiment lexicon DUTIR [39] for Chinese reviews. The DUTIR lexicon contains 27446 common Chinese words. The sentiment polarity of these words is labeled as positive, negative, or neutral.

(DUTIR) We consider only the words contained in the DUTIR lexicon as features, as in the case of the bag-of-words model. Therefore, the review d can be represented as $\overrightarrow{d_0} = [0, 1, \ldots, 1]$.

(Only 3) We consider the three polarities of the sentiment words in the DUTIR lexicon. We represent the product review d by a three-dimensional vector $\overrightarrow{d_1} = [k_0, k_1, k_2]$, where k_0, k_1, and k_2 are the number of words with the three types of polarities in the review.

(DUTIR+3) Here, we combine the above two representations. The product review d can be represented as $\overrightarrow{d} = \overrightarrow{d_0} + \overrightarrow{d_1} = [0, 1, \ldots, 1, k_0, k_1, k_2]$.

For the English reviews, we choose a general sentiment lexicon SentiWordNet (http://sentiwordnet.isti.cnr.it/). SentiWordNet assigns three sentiment scores to each synset of WordNet: positivity, negativity, and objectivity. In other words, both the sentiment dimensions of the words for these two general sentiment lexicons in Chinese and English are 3.

(SentiWordNet) As in the case of the DUTIR lexicon, we focus on the words contained in the SentiWordNet lexicon. The review d can be represented as $\overrightarrow{d_0} = [0, 1, \ldots, 1]$.

(Only 3) We represent the product review d by a three-dimensional vector $\overrightarrow{d_1} = [k_0, k_1, k_2]$. Here, k_0 is the sum of the positivity scores of the words in the review d. Similarly, k_1 and k_2 represent the sum of the negativity and objectivity scores, respectively.

(SentiWordNet+3) As in the case of the DUTIR lexicon, the product review d can be represented as $\overrightarrow{d} = \overrightarrow{d_0} + \overrightarrow{d_1} = [0, 1, \ldots, 1, k_0, k_1, k_2]$.

(c) Domain-Specific Lexicon

We use hybrid approaches to evaluate the domain-specific lexicon developed using our method. We set the window size as 3 and α as 0.01 (as mentioned in Section 3.2). We use k-means (http://scikit-learn.org/stable/modules/clustering.html#k-means) to cluster the sentiment words into k categories based on the distance in the matrix space. Unlike the general lexicons, the sentiment dimension of words in our domain-specific lexicon is k. Note that we select k through a fivefold cross-validation on the training set. The details of the selection of k are explained in the next subsection. In our experiment, the number

of clusters is not more than 30. In the following discussion, the domain-specific lexicon built using our method is denoted as DS.

(DS) We consider the sentiment words contained in the domain-specific lexicon as features, as in the case of the bag-of-words model. Accordingly, the review d can be represented as $\overrightarrow{d_0} = [0, 1, \ldots, 1]$.

(Only *k*) We cluster the sentiment words into k categories using k-means. We represent the product review d by a k ($k \geq 2$)-dimensional vector $\overrightarrow{d_1} = [m_0, m_1, .., m_{k-1}]$. Obviously, m_t represents the number of sentiment words that belong to the (t + 1)th category in the review.

(DS+*k*) Here, we combine the above two representations. The product review d can be represented as $\overrightarrow{d} = \overrightarrow{d_0} + \overrightarrow{d_1} = [0, 1, \ldots, 1, m_0, m_1, .., m_{k-1}]$.

Three different sentiment matrices are considered in our lexicon construction process. The sentiment word representations in different matrices are very different. Therefore, the clustering result would also be different. We use k(PMI), k(TFIDF-PMI), and k(EPMI) to represent the clustering results of sentiment matrices A, B, and F, respectively. We discuss the different results of the three matrices in the next subsection.

In addition, we evaluate the lexicon in terms of the coverage and usage. We assume that the test set contains N unique words and that these N words include T sentiment words, which are contained in the sentiment lexicon. We also assume that the size of the lexicon that is used to train the classification model is S. Therefore, the coverage of the lexicon is T/N, and the usage of lexicon is T/S. If the coverage of the lexicon is low, the classification performance will be unsatisfactory. If the usage of the lexicon is low, computing resources will be wasted, which should be avoided especially for mobile devices. Considering these two evaluation indexes, we propose an average evaluation index such as the F1-measure. Let V, U, and T represent the coverage, usage, and average of the lexicon, respectively. Then, T is defined as

$$\mathrm{T} = \frac{2 * \mathrm{V} * \mathrm{U}}{\mathrm{V} + \mathrm{U}} * 100\% \tag{9}$$

4.3. Results and Discussion

Overall Results. Table 5 lists the overall classification results. All the tasks are balanced two-class problems. The best result for each domain review is marked in bold font, and the second-best result is underlined.

First, for the domain-specific and general lexicons, DS+k achieves the best results for all the seven domain reviews whereas DS achieves the second-best results for four of the reviews. DS outperforms the general lexicons DUTIR and SentiWordNet. These results indicate that the domain-specific lexicon, which is constructed from the corresponding corpus, shows better performance for sentiment classification tasks on shopping reviews.

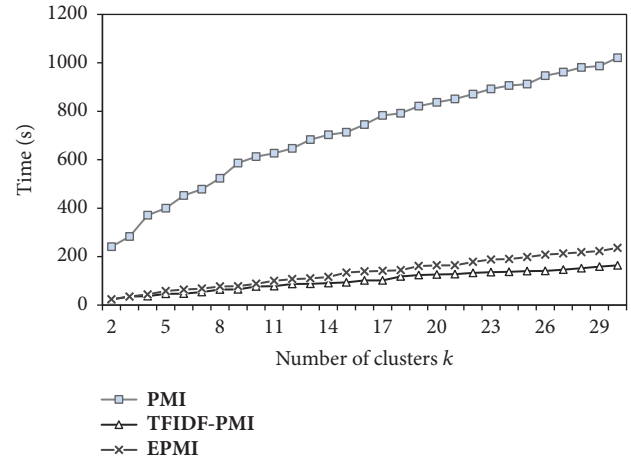


Figure 2: Time required for clustering in sentiment matrix built using three methods. The three different lines represent three different sentiment matrices.

Second, for no lexicon approaches, the classical bag-of-words model obviously performs better than the deep-learning model Word2Vec in terms of sentiment classification tasks. BOW achieves the second-best results for three of the reviews, whereas W2V shows nearly the worst performance for both Chinese and English reviews. The poor performance is unexpected, and a large corpus of training data is perhaps required for training Word2Vec [40].

Third, for sentiment dimensions, the performances of Only 3 and Only k are relatively worse for both Chinese and English reviews. That is, it is not sufficient to just consider the sentiment dimensions when we use the lexicon as the source of the features to express the reviews. However, the performance of DS+k is better than those of DS and Only k for both Chinese and English reviews. This result indicates the effectiveness of combining the words and sentiment dimensions of the lexicon.

Note that, in Table 5, k represents k(EPMI). Considering that (DS+k) provides the best performance, we analyze the differences among the results of DS+k(PMI), DS+k(TFIDF-PMI), and DS+k(EPMI) in detail.

EPMI versus PMI and TFIDF-PMI. Table 6 lists the classification results of (DS+k) with the three different methods mentioned in Section 3. First, we find that the classification performance of DS+k(EPMI) is better than those of DS+k(PMI) and DS+k(TFIDF-PMI). In particular, the performance of DS+k(PMI) is relatively poor. According to a t-test, the differences among the results of the three methods are significant ($p < 0.05$). This result reflects the advantages of EPMI over traditional PMI in sentiment classification.

Second, we discuss the differences among the three methods in terms of time efficiency. Figure 2 shows the average clustering times required by the three different sentiment

Table 5: F1-measure classification results for shopping domain reviews (NB).

Method	Chinese			English			
	Hotel	Cloth	Fruit	Books	Kitchen	DVD	Electronics
No lexicon							
BOW	81.9	86.4	88.8	74.0	79.3	77.6	77.0
W2V	61.2	68.2	69.8	61.5	67.1	63.8	61.7
General lexicon							
DUTIR	71.5	75.0	73.8	-	-	-	-
SentiWordNet	-	-	-	72.3	79.0	76.8	76.5
Only 3	68.8	70.7	70.2	65.1	62.3	68.4	61.9
DUTIR+3	72.6	74.8	74.2	-	-	-	-
SentiWordNet+3	-	-	-	74.9	79.8	78.5	77.9
Domain-specific lexicon							
DS	80.3	84.9	86.8	77.7	81.1	79.5	79.0
Only k	70.3	71.2	71.5	61.6	64.9	62.7	63.2
DS+k	**83.6**	**87.6**	**89.5**	**78.9**	**82.5**	**80.6**	**80.5**

Table 6: Classification results of DS+k with three different methods.

DS + k	Chinese			English			
	Hotel	Cloth	Fruit	Books	Kitchen	DVD	Electronics
k(PMI)	81.0	85.2	87.3	77.3	81.5	79.0	79.2
k(TFIDF-PMI)	82.8	87.1	89.0	78.4	82.1	80.1	79.9
k(EPMI)	**83.6**	**87.6**	**89.5**	**78.9**	**82.5**	**80.6**	**80.5**

matrices. The time consumed by sentiment matrices B and F is considerably less than that consumed by sentiment matrix A. This is because in matrices B and F, the dimension of the vector in the matrix space is reduced by using the method described in Section 3.2. The dimension reduction leads to a substantial increase in the efficiency and accuracy of classification. Therefore, sentiment matrix F constructed using EPMI shows the best performance.

Selection of K. The sentiment dimensions of the domain-specific lexicon is k. Now, we analyze the influence of different k values on the classification performance. Figure 3 shows the performance of Only k(EPMI) with the change in k for the English product reviews. When k is 2, Only k(EPMI) shows the best performance for the books and DVD domains. For the kitchen and electronics domains, a larger k improves the classification performance of Only k(EPMI). The appropriate value of k for the domain-specific lexicon is different for different fields. We select the value of k through a fivefold cross-validation on the training set in our experiments.

However, we find that the performance of Only k(EPMI) is worse than that of Only 3 in the books and DVD domains (Table 5). We believe that this is because $k = 2$ is not a good choice for Only k(EPMI) for performing sentiment classification tasks. To prove this point, we look at the performance results of Only $k(k = 2)$ for all English product reviews (Table 7).

Table 7 indicates that the performance of Only k(EPMI) is not good when k is fixed at 2. In our domain-specific lexicon, low sentiment dimensions such as $k = 2$ is not good for DS. k has a substantial influence on our domain-specific lexicon for sentiment classification tasks. Therefore, it is necessary to select the k value by cross-validation.

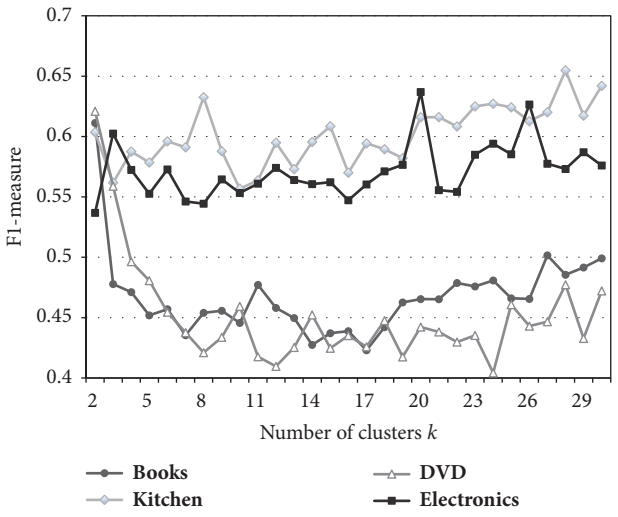


Figure 3: Performance of Only k(EPMI) with change in k.

Table 7: Performance results of Only k (EPMI). k is selected using two different methods.

	Selection of k	Books	Kitchen	DVD	Electronics
Only k	$k = 2$	61.1	60.3	62.1	53.6
	Cross-validation	61.6	64.9	62.7	63.2

Table 8: F1-measure classification results for shopping domain reviews (SVM).

Methods	Chinese			English			
	Hotel	Cloth	Fruit	Books	Kitchen	DVD	Electronics
No lexicon							
BOW	73.5↓	76.1↓	82.1↓	75.8↑	79.4	78.0↑	77.2
General lexicon							
DUTIR+3	79.5↑	82.1↑	82.5↑	-	-	-	-
SentiWordNet+3	-	-	-	74.1↓	76.2↓	74.9↓	76.3↓
Domain-specific lexicon							
DS+k	82.3↓	84.2↓	86.8↓	79.1	80.9↓	79.2↓	78.3↓

Table 9: Coverage, usage, and average of lexicons for Chinese reviews.

	BOW			DUTIR			DS		
	V	U	T	V	U	T	V	U	T
Hotel	71.0	34.1	46.1	7.7	2.5	3.8	24.5	39.6	30.2
Cloth	73.4	36.9	79.1	8.2	2.7	4.1	29.0	40.2	33.7
Fruit	75.0	37.5	50.0	8.0	3.3	4.0	28.9	39.9	31.9

Table 10: Coverage, usage, and average of lexicons for English reviews.

	BOW			SentiWordNet			DS		
	V	U	T	V	U	T	V	U	T
Books	62.3	24.1	34.7	51.6	5.2	9.4	37.9	32.5	34.9
Kitchen	67.1	28.2	39.7	54.8	2.7	5.1	42.8	39.5	41.1
DVD	60.3	24.2	34.5	46.6	5.0	9.0	34.2	33.2	33.6
Electronics	63.3	25.9	36.7	49.6	2.8	5.2	38.5	35.1	36.7

NB versus SVM. For obtaining all the above results, we have chosen NB as the classifier. However, classification algorithms influence the classification performance. Therefore, we choose another popular classification algorithm SVM as the classifier for the method with the best performance among each type of methods. The results are listed in Table 8, where ↑ indicates an improvement in performance compared to that when NB is used and ↓ indicates a deterioration in performance. SVM performs better than NB when using DUTIR+3 for Chinese reviews and when using BOW for the books and DVD domains. In contrast, NB yields better performance when using the other approaches. Sentiment classification is perhaps one of the domains that have clear feature dependence, and hence, NB often performs unexpectedly well [41]. Although the domain-specific lexicon performs better with both NB and SVM, different types of models of text classification are probably required for documents with different properties. Hence, further empirical and theoretical study is required to understand the relationship between sentiment classification tasks and classification models.

Lexicon Coverage. Finally, we discuss the classification performance in terms of the coverage (V), usage (U), and average (T). The results for the test set are listed in Tables 9 and 10. In both the Chinese and English domains, the average of BOW is relatively high. For Chinese product reviews, both the coverage and usage of DUTIR are the worst because DUTIR is a general lexicon that contains only few words that often appear in shopping reviews. The coverage of SentiWordNet is considerably higher than that of DUTIR. This partly explains why the performance of SentiWordNet is better than that of DUTIR for sentiment classification tasks. The better performance is also probably because SentiWordNet contains more words related to mobile shopping reviews than DUTIR. The coverage of SentiWordNet is higher than that of the domain-specific lexicon for English product reviews, whereas the usage of SentiWordNet is considerably low than that of DS. The very low usage of lexicons may impact their performance and waste the computing resources of mobile devices. The average of DS is considerably higher than that of the general lexicon for both Chinese and English product reviews. This result reflects the advantage of our domain-specific lexicon for mobile shopping reviews in another way.

5. Conclusions

The analysis of the sentiment of users' product reviews largely depends on the quality of sentiment lexicons. This paper presents a sentiment lexicon construction approach for mobile shopping. In this approach, a sentiment matrix that considers the relationship between sentiment words and product features is built. The sentiment words are clustered based on the distance between them in the matrix space. One characteristic of our lexicon is that the sentiment words are clustered into several categories rather than into binary or other fixed categories. In other words, the sentiment

dimension of the words in our lexicon is flexible. In addition, the product features are filtered based on the idea of TFIDF. Moreover, the EPMI algorithm is proposed, which is more appropriate for the mobile review domain. The experimental results show that our sentiment lexicons outperform the benchmarks with statistically significant differences in terms of sentiment classification tasks, thus proving the effectiveness of the proposed approach.

Data Availability

The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The work described in this paper was partially supported by the National Key R&D Program of China (Grant no. 2018YFC0407901), partially supported by the National Natural Science Foundation of China under Grants no. 61370091 and no. 61602149, partially supported by the Fundamental Research Funds for the Central Universities under Grant no. 2016B01714, and partially supported by the Priority Academic Program Development of Jiangsu Higher Education Institutions. Lau's work is supported by grants from the Research Grants Council of the Hong Kong SAR (Projects CityU 11502115 and CityU 11525716) and the NSFC Basic Research Program (Project 71671155).

References

[1] A. Montoyo, P. Martínez-Barco, and A. Balahur, "Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments," *Decision Support Systems*, vol. 53, no. 4, pp. 675–679, 2012.

[2] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing—Volume 10 (EMNLP '02)*, pp. 79–86, Association for Computational Linguistics, Stroudsburg, Pa, USA, July 2002.

[3] P. D. Turney, "Thumbs up or thumbs down?" in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 417–424, Philadelphia, Pennsylvania, July 2002.

[4] A. Ortigosa, J. M. Martín, and R. M. Carro, "Sentiment analysis in Facebook and its application to e-learning," *Computers in Human Behavior*, vol. 31, no. 1, pp. 527–541, 2014.

[5] B. Liu, "Sentiment analysis and opinion mining," *Morgan & Claypool*, 2012.

[6] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04)*, pp. 168–177, August 2004.

[7] Y.-J. Tai and H.-Y. Kao, "Automatic domain-specific sentiment lexicon generation with label propagation," in *Proceedings of the International Conference on Information Integration and Web-Based Applications Services*, pp. 53–62, 2013.

[8] D. Maynard and A. Funk, "Automatic detection of political opinions in tweets," in *Proceedings of the 1st Workshop on Making Sense of Microposts 2011: Big Things Come in Small Packages, MSM 2011 - Co-located with the 8th Extended Semantic Web Conference, ESWC 2011*, pp. 81–92, Greece, May 2011.

[9] Z. Zhang, Q. Ye, Z. Zhang, and Y. Li, "Sentiment classification of internet restaurant reviews written in cantonese," *Expert Systems with Applications*, vol. 38, no. 6, pp. 7674–7682, 2011.

[10] X. Li, J. Cao, and Z. Pan, "Market impact analysis via deep learned architectures," *Neural Computing and Applications*, pp. 1–12, 2018.

[11] F. Enríquez, J. A. Troyano, and T. López-Solaz, "An approach to the use of word embeddings in an opinion classification task," *Expert Systems with Applications*, vol. 66, pp. 1–6, 2016.

[12] Z. Hailong, G. Wenyan, and J. Bo, "Machine learning and lexicon based methods for sentiment classification: A survey," in *Proceedings of the 11th Web Information System and Application Conference, WISA 2014*, pp. 262–265, China, September 2014.

[13] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in *Proceedings of the 2nd International Conference on Knowledge Capture, K-CAP 2003*, pp. 70–77, USA, October 2003.

[14] G. Qiu, X. He, F. Zhang, Y. Shi, J. Bu, and C. Chen, "DASA: Dissatisfaction-oriented Advertising based on Sentiment Analysis," *Expert Systems with Applications*, vol. 37, no. 9, pp. 6182–6191, 2010.

[15] S. Jadav, B. Tanawal, and H. Guadani, "Sentiment analysis: a review," *International Journal of Advance Engineering and Research Development*, vol. 4, pp. 957–962, 2017.

[16] V. Sindhwani and P. Melville, "Document-word co-regularization for semi-supervised sentiment analysis," in *Proceedings of the 8th IEEE International Conference on Data Mining, ICDM 2008*, pp. 1025–1030, Italy, December 2008.

[17] X. Li, H. Xie, L. Chen, J. Wang, and X. Deng, "News impact on stock price return via sentiment analysis," *Knowledge-Based Systems*, vol. 69, no. 1, pp. 14–23, 2014.

[18] G. Badaro, R. Baly, R. Akel et al., "A Light Lexicon-based Mobile Application for Sentiment Mining of Arabic Tweets," in *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pp. 18–25, Beijing, China, July 2015.

[19] J. Kamps, M. Marx, R. J. Mokken, and M. De Rijke, "Using WordNet to measure semantic orientations of adjectives," in *Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004*, pp. 1115–1118, Portugal, May 2004.

[20] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: a survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[21] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics (ACL '97)*, pp. 174–181, 1997.

[22] P. D. Turney and M. L. Littman, "Measuring praise and criticism: inference of semantic orientation from association," *ACM Transactions on Information and System Security*, vol. 21, no. 4, pp. 315–346, 2003.

[23] A. Yang, J. Lin, Y. Zhou, and J. Chen, "Research on building a Chinese sentiment lexicon based on SO-PMI," *Applied Mechanics and Materials*, vol. 263-266, no. 1, pp. 1688–1693, 2013.

[24] P. Ekman, "An Argument for Basic Emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[25] V. L. Rubin, J. M. Stanton, and E. D. Liddy, *Discerning Emotions in Texts*, Stanford University, 2004.

[26] A. Tellegen, D. Watson, and L. A. Clark, "On the dimensional and hierarchical structure of affect," *Psychological Science*, vol. 10, no. 4, pp. 297–303, 1999.

[27] E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall, "Sentiment Analysis Is a Big Suitcase," *IEEE Intelligent Systems*, vol. 32, no. 6, pp. 74–80, 2017.

[28] E. Fast, B. Chen, and M. S. Bernstein, "Empath: Understanding topic signals in large-scale text," in *Proceedings of the 34th Annual Conference on Human Factors in Computing Systems, CHI 2016*, pp. 4647–4657, USA, May 2016.

[29] H. Zhang, Z. Yu, M. Xu, and Y. Shi, "Feature-level sentiment analysis for Chinese product reviews," in *Proceedings of the 3rd International Conference on Computer Research and Development (ICCRD)*, pp. 135–140, Shanghai, China, March 2011.

[30] M. Gilad, "Experiments with mood classification in blog posts," *ACM Transactions on Multimedia Computing Communications and Applications*, 2005.

[31] F. Benamara, C. Cesarano, A. Picariello, D. Reforgiato, and V. S. Subrahmanian, "Sentiment analysis: Adjectives and adverbs are better than adjectives alone," in *Proceedings of the 2007 International Conference on Weblogs and Social Media, ICWSM 2007*, USA, March 2007.

[32] Y. Lu, X. Kong, X. Quan, W. Liu, and Y. Xu, "Exploring the Sentiment Strength of User Reviews," in *Web-Age Information Management*, vol. 6184 of *Lecture Notes in Computer Science*, pp. 471–482, Springer, Berlin, Germany, 2010.

[33] T. Wilson, P. Hoffmann, S. Somasundaran et al., "OpinionFinder," in *Proceedings of the HLT/EMNLP*, pp. 34-35, Vancouver, British Columbia, Canada, October 2005.

[34] C. Hou, F. Nie, D. Yi, and D. Tao, "Discriminative embedded clustering: a framework for grouping high-dimensional data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 6, pp. 1287–1299, 2015.

[35] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL '07)*, pp. 440–447, June 2007.

[36] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS '13)*, pp. 3111–3119, December 2013.

[37] F. Pedregosa, G. Varoquaux, and A. Gramfort, "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[38] D. Zhang, H. Xu, Z. Su, and Y. Xu, "Chinese comments sentiment classification based on word2vec and SVMperf," *Expert Systems with Applications*, vol. 42, no. 4, pp. 1857–1863, 2015.

[39] H. Lin, L. Xu, H. Ren et al., "Constructing the affective lexicon ontology," *Journal of The China Society for Science and Technical Information*, vol. 27, pp. 180–185, 2008.

[40] A. Edgar, R. Sidarta, S. Mariano, and F. S. Diego, "Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database," in *ASAI Simposio Argentino de Inteligencia Artificial*, 2016.

[41] P. Domingos and M. Pazzani, "Beyond independence: Conditions for the optimality of the simple bayesian classifier," *Machine Learning*, vol. 29, pp. 103–130, 1996.

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 2678976, 8 pages
https://doi.org/10.1155/2018/2678976

Research Article

Using Sentence-Level Neural Network Models for Multiple-Choice Reading Comprehension Tasks

Yuanlong Wang,[1] **Ru Li**,[1,2] **Hu Zhang**,[1] **Hongyan Tan**,[1] **and Qinghua Chai**[3]

[1]*School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China*
[2]*Key Laboratory of Computation Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, China*
[3]*School of Foreign Languages, Shanxi University, Taiyuan 030006, China*

Correspondence should be addressed to Yuanlong Wang; ylwang@sxu.edu.cn

Received 28 March 2018; Accepted 13 June 2018; Published 3 July 2018

Academic Editor: Tianyong Hao

Comprehending unstructured text is a challenging task for machines because it involves understanding texts and answering questions. In this paper, we study the multiple-choice task for reading comprehension based on MC Test datasets and Chinese reading comprehension datasets, among which Chinese reading comprehension datasets which are built by ourselves. Observing the above-mentioned training sets, we find that "sentence comprehension" is more important than "word comprehension" in multiple-choice task, and therefore we propose sentence-level neural network models. Our model firstly uses LSTM network and a composition model to learn compositional vector representation for sentences and then trains a sentence-level attention model for obtaining the sentence-level attention between the sentence embedding in documents and the optional sentences embedding by dot product. Finally, a consensus attention is gained by merging individual attention with the merging function. Experimental results show that our model outperforms various state-of-the-art baselines significantly for both the multiple-choice reading comprehension datasets.

1. Introduction

Reading comprehension is the ability of reading texts, understanding their meanings, and answering questions. When machines are required to comprehend texts, they need to understand unstructured text and do reasoning based on the text [1–3]. It is a major task in the field of natural language processing and machine learning.

Recently, machine reading comprehension (MC) is increasingly drawing attention and several large reading comprehension datasets have also been released. For the several released datasets, the task is getting more and more difficult (from CNN/Daily Mail datasets to SQuAD and then to TriviaQA) as system performance has rapidly improved with each new released datasets. The CNN/Daily Mail datasets [4] **is a cloze-style reading comprehension task,** which aims to comprehend a given document and then to answer questions based on the given document, and the answer to each question is a single word inside of the document. The SQuAD [5] **is a question-answering reading comprehension task**, which further constrains answers often including nonentities and being much longer phrases to be a continuous subspan of the document. Clearly, the question-answering task is more difficult than the cloze-style task. The TriviaQA [6] **is also a question-answering reading comprehension task,** but the task in TriviaQA is more difficult than the task in SQuAD because answers in TriviaQA are independent of the evidence and belong to a diverse set of types.

Different from the above, the task based on the MCTest datasets [3] **is a multiple-choice reading comprehension,** each example of which consists of one document and four associated questions and each question gives four candidate answers and only one answer is correct among them. In this paper, we focus on such problem of answering multiple-choice questions in documents, and, at the same time, we also release a Chinese reading comprehension dataset for such multiple-choice task. To our knowledge, the dataset is the first Chinese reading comprehension dataset of this

Document:
"Ruins" is a derogatory term that it is irrelevant to cultural and aesthetic in many Chinese mind, and interpretation of the word "ruins" is only a "city and village are changed into desolate places by destruction or natural disasters" in the "Modern Chinese Dictionary"; There is no fault for the interpretation, but it is not enough if it is measured by world knowledge. In Europe, the meaning of "ruins" has been enriched and expanded since modern times. It has been endowed with the connotation of culture and aesthetics, and has become an academic concept. The of meaning of the "ruins" is changed from the Renaissance in Europe.
Question:
Please choice two incorrect options according to the content of the document:
Choice:
A. One of the purposes of this paper is to correct the misunderstanding of the term "ruins" in the modern Chinese dictionary.
B. The Great Wall Ruins have condensed the vicissitudes of time in China and it have a "perception of the intoxicated" as the Acropolis ruins.
C. Remains of the ruins often reveals the extraordinary wisdom and great efforts of the predecessors, which bring to the future generations with the shock and resonance of the soul.
D. Awareness of the ruins is related to the aesthetic consciousness of countrymen, but also it is conducive to the popularity of the "repair the old as the old".
E. This paper not only contains historical interest, but also infiltrated the concern of reality, and express the author's desire to enhance the cultural quality of the nation.

Box 1: Example for the multiple-choice reading comprehension for literature (the original data is in Chinese, we translate this sample in English for clarity).

kind and is even more complex than MCTest datasets. The example of such dataset consists of one document and one associated question which gives five candidate answers. The specific details of this dataset are in Section 2. Frankly, the **multiple-choice reading comprehension** task remains quite challenging. For one thing, answers in the form of an optional sentence usually do not appear in the document; for another, finding the correct answer of the given question requires reasoning across multiple sentences. Hence, sentence comprehension is more important than word comprehension in the task of the multiple-choice reading comprehension.

To carry out the task of sentence comprehension, we propose a sentence-level attention model primarily inspired by the attention model for the Cloze-style reading comprehension [7, 8]. However, unlike the Cloze-style attention model, answers to multiple-choice questions are optional sentences. Karl et al. [9] train an encoder-decoder model to encode a sentence into a fixed length vector and subsequently decode both the following sentences. They also demonstrate that the low-dimensional vector embeddings are useful for other tasks. Pichotta et al. [10] present a sentence-level LSTM language model for script inference. The results show that the model is useful for predicting missing information in text. Similar to the above model, we also present a sentence representation model which uses LSTM network to learn vector representation for sentences. Moreover, we use sentence composition model to represent sentence vector because the model can express hierarchical sentences from words to phrases, and to sentences. In order to retain more information about two kinds of sentences representation model, we employ connection method to compose the final sentence vector. Then, we train a sentence attention model between optional sentences and sentences in the document. The machine is able to learn the relationships between the document and optional sentences by the attention-based neural network.

Experimental results show that our approach can effectively improve the performance of the task of multiple-choice reading comprehension. In the following text, Chinese reading comprehension datasets, related work, details of our model, and experiments will be described, and, afterwards, our experiments will be analyzed.

2. Chinese Reading Comprehension Datasets

In this paper, we focus on the multiple-choice reading comprehension task. Similar to the MCTest datasets, each example consists of one document and one associated questions. And each question gives five candidate answers. However, the dataset is more complex than MCTest datasets, and it is a literary reading comprehension dataset from test materials of final exam in senior high school. Box 1 shows an example of Chinese reading comprehension datasets.

For the dataset, the description of questions is basically fixed, as in the following: "Question". Therefore, the role of question is ignored in the Chinese reading comprehension task. The goal of the task is to understand the individual document and to select the most consistent options with the meaning of the document. Thus the Chinese reading comprehension can be described as a triple:

$$\langle D, C, A \rangle \tag{1}$$

where D is the document, C denotes the choice, and A is a set in which each element is marked as 0 or 1 according to the document meaning (if the option is consistent with the document meaning, it is labeled as 1; otherwise it is labeled as 0). The A can be described as the following:

Question: "Please choose two incorrect options according to the content of the document: "

Answer: C E

A(Answer label): (1 1 0 1 0)

Table 1: Statistics of multiple-choice reading comprehension datasets: train and three tests.

	Documents	Sentences	options
Train	769	28235	3845
BCEETest	13	548	65
SBCEETest1	12	517	60
SBCEETest2	52	2056	260

In the training stage, we choose a 769-literary-reading-comprehension dataset which is collected from test materials of final exam in senior high school. In the testing stage, the dataset includes three parts: 13 Beijing college entrance examination papers (BCEETest), 12 simulation materials (SBCEETest1) which is provided by iFLYTEK company, and 52 test materials of final exam in Beijing senior high school (SBCEETest2). All of datasets are collected by the Chinese information processing group of Shanxi University. The statistics of training and testing data are shown in Table 1.

3. Related Work

Machine comprehension is currently a hot topic within the machine learning community. In this section we will focus on the best-performing models applied to MCTest and CNN/Daily Mail according to two kinds of reading comprehension tasks.

3.1. Multiple-Choice Reading Comprehension. Existing models are mostly based on manually engineered features for MCTest [11–13]. These engineered feature models are extremely effective. However, this research often requires significant effort on the auxiliary tools to extract the feature and its capacity for generalization is limited.

Yin et al.[14] proposed a hierarchical attention-based convolutional neural network for multiple-choice reading comprehension task. The model considers multiple levels of granularity, from word to sentence level and then from sentence to snippet level. This model performs poorly on MCTest. A possible reason that can explain this is that the dataset is sparse. However, neural model can address the extracted features problem, so it appeals to increasing interest in multiple-choice reading comprehension task. For sequence data, the recurrent neural network often is used. So we propose a recurrent neural network model for the multiple-choice reading comprehension. Our model uses the bidirectional LSTM to get contextual representations of the sentence.

3.2. Cloze-Style Reading Comprehension. Hermann et al. [4] published the CNN/Daily Mail news corpus, where the content is formed by news articles and its summarization. Also, Cui et al. [7] released HFL-RC PD&CFT for Chinese reading comprehension datasets, which includes People Daily news datasets and Children 's Fairy tale datasets. On these datasets, many neural network models have been proposed for the Cloze-style reading comprehension tasks. Hermann et al. [4] proposed the attentive and impatient readers. The attentive reader uses bidirectional document and query encoders to compute an attention and the impatient reader computes attention over the document after reading every word of the query. Chen et al. [1] proposed a new neural network architecture for the Cloze-style reading comprehension. In contrast to the attentive reader, the attention weights of the model are computed with a bilinear term instead of simple dot product. Kadlec et al. [15] proposed the Attention Sum Reader, which uses attention to directly pick the answer from the context. The model uses attention as pointer over discrete tokens in the context document and then directly sums the word attention across all the occurrences. Cui et al. [7] presented the consensus attention-based neural network, namely, Consensus Attention Sum Reader, and released Chinese reading comprehension datasets. The model computes an attention to every time slice of query and makes a consensus attention among different steps. Cui et al. [8] also proposed the attention-over-attention neural network, namely, Attention-over-Attention Reader. The model presents an attention mechanism that places another attention over the primary attention, to indicate the "importance" of each attention. Dhingra et al.[16] proposed the gated-attention readers for text comprehension. The model integrates a multihop architecture with an attention mechanism which is based on multiplicative interactions between the query embedding and the intermediate states of a recurrent neural network document reader.

To summarize, all of them are attention-based RNN models which have been shown to be extremely effective for the word-level task. At each time-step, these models take a word as input, update a hidden state vector, and predict the answer. In this paper, we propose sentence-level attention model for the multiple-choice reading comprehension. Our work is primarily inspired by the attention model for the Cloze-style reading comprehension.

4. Sentence-Level Neural Network Reader

In this section, we will introduce our sentence-level neural network models for the multiple-choice reading comprehension task, namely, Sentence-Level Attention Reader. Our model is primarily motivated by that of Cui et al. [7], which aims to directly estimate the answer of optional sentence from the sentence-level attention instead of calculating the answer of entity from the word-level attention. The level structure of our model is shown in Figure 1. Firstly, the document is divided into several sentences $D = \{s_1, s_2, \ldots s_n\}$ and the sentence embedding is computed by embedding layer. Secondly, we use the bidirectional LSTM to get contextual representations of the sentence, in which the representation of each sentence is formed by concatenating the forward and backward hidden states. Thirdly, the sentence-level attention is computed by a dot product between the sentence embedding in the document and the optional embedding. Finally, the individual attention is merged to a consensus attention by the merging function. The following will give a formal description of our proposed model.

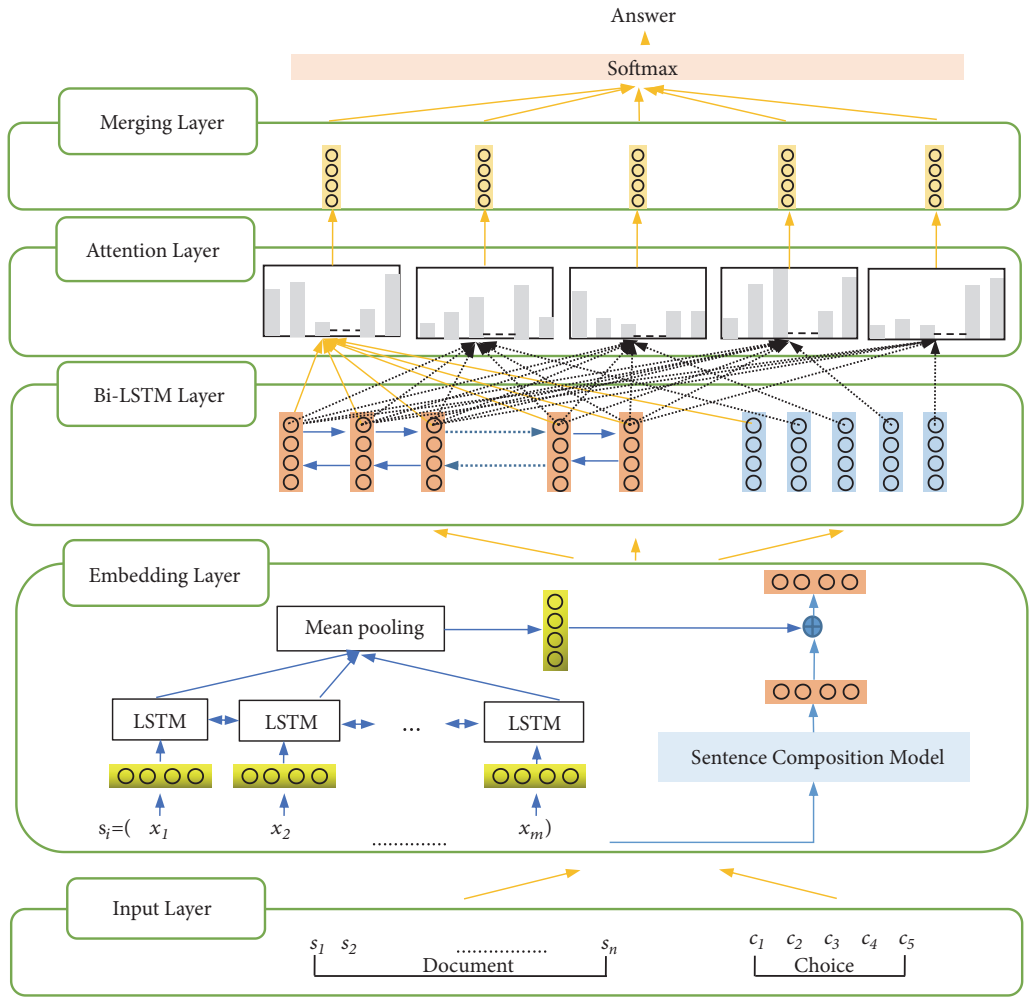


FIGURE 1: Sentence-level attention neural network. The network includes 5 layers: input layer, embedding layer, Bi-LSTM layer, attention layer, and merging layer.

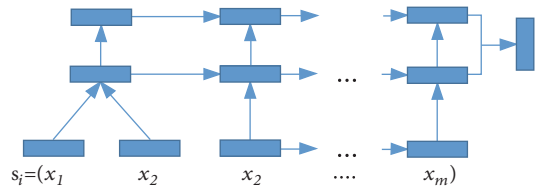


FIGURE 2: Sentence composition model.

4.1. Sentence Representation. The input of our model is the sentences in the document and options, and each sentence consists of word sequence. The sentence is translated into sentence embedding by embedding layer, which is composed of LSTM sentence model and sentence composition model[17] as illustrated in the embedding layer of Figure 1. The LSTM sentence model is a single bi-LSTM layer followed by an average pooling layer. The bi-LSTM layer is used to get the contextual representations of words and the average pooling layer is used to merge word vectors into sentence vectors. On the other hand, we used the sentence composition model to compose sentence vector. The sentence vector is combined by the trained neural network model, which is trained by the triple consisting of single words and phrases vector (as triple(w_1, w_2, p)). The sentence composition model is illustrated in Figure 2. We denote p_i as the final sentence vector. In order to retain more information about two kinds of sentences' representation model, we employ a multi-layer neural network to compose the final sentence vector, $p_i(sp_1, sp_2) = sp_1{}^T M\, sp_2$, where sp_1 is the sentence vector for LSTM sentence model, sp_2 is the sentence vector for sentence composition model, and M is a parameter matrix.

In addition to the representation of sentences mentioned above, the context of sentence is also important for inferring

Table 2: Experimental results for MCTest.

Method	MC160 Test	MC500 Test
Richardson et al.(2013)+RTE	0.691	0.633
Sachan et al.(2015)	-	0.678
Wang et al.(2015)	0.753	0.699
Trischler et al.(2016)	0.746	0.710
Attentive Reader	0.463	0.419
Neural Reasoner	0.476	0.456
HABCNN-TE	0.631	0.529
Sentence-Level Attention Reader (mode:max+avg)	0.664	0.673

the answer. So the embedding of the sentence in the document is inputted into bi-LSTM layer to get their contextual representations. In our model, the bidirectional LSTM is used as RNN implementation.

$$\overrightarrow{h_i} = LSTM\left(\overrightarrow{h_{i-1}}, p_i\right), \quad i = 1, 2, \ldots m \tag{2}$$

$$\overleftarrow{h_i} = LSTM\left(\overleftarrow{h_{i+1}}, p_i\right), \quad i = m, m-1, \ldots 1 \tag{3}$$

$$h_i = concat\left(\overrightarrow{h_i}, \overleftarrow{h_i}\right) \tag{4}$$

Finally, we take h_i to represent the contextual representations of sentences. $h_{c\_s} \in R^d$ denote the sentence embedding of the option, where d denotes the number of options.

4.2. Sentence-Level Attention. In attention layer, we directly use a dot product of h_i and h_{c_s} to compute the "importance" of each sentence in the document for each option. And we use the softmax function to get a probability distribution. For each sentence in the document, "attention" is computed as follows.

$$\alpha(t) = softmax\left(h_i(t) \bullet h_{c\_s}\right) \tag{5}$$

where variable $\alpha(t)$ is the attention weight a tth sentence in document.

In merging layer, the consensus attention is calculated by a merging function as follows.

$$s \propto \begin{cases} \sum_{t=1}^{m} \alpha(t), & if\ mode = sum; \\ \frac{1}{m} \sum_{t=1}^{m} \alpha(t), & if\ mode = avg; \\ \max_{t=1\ldots m} \alpha(t), & if\ mode = max; \\ \frac{1}{n} \sum_{t=1}^{n} \alpha(t), & if\ mode = max + avg, \end{cases} \tag{6}$$

where n is the top number of the attention weight and $n < m$.

4.3. Output Layer. Finally, the answer is estimated by the softmax function.

$$a_i = soft\ \max\left(W_a * s_i\right), \quad i = 1 \ldots 5 \tag{7}$$

where W_a indicate the weight matrix in the softmax layer and a_i is a probability distribution of the answer. The prediction of answer labels (such as "1 1 0 1 0") is gotten by the probability. Figure 1 shows the proposed neural network architecture.

5. Experiments

In this section we evaluate our model on the MCTest and our Chinese reading comprehension datasets. We find that although the model is simple, it achieves state-of-the-art performance on these datasets.

5.1. Experimental Details. We use stochastic gradient descent with AdaDelta update rule [18], which only uses the first-order information to adaptively update learning rate over time and has minimal computational overhead. To train model, we minimize the negative log-likelihood as the objective function. The batch size is set to 5 and the number of iterations is set to 25.

For word vectors we use Google's publicly available embedding [19], whose training dataset is 70 thousand literary papers. The dimension of word embedding is set to 200. While we are implementing the sentence-level attention reader, it is easy to overfit the training data. Thus, we adopt dropout method [20] for regularization purpose and handling overfitting problems. The dropout rate is 0.1 on Chinese reading comprehension datasets and 0.01 on MCTest datasets, respectively. Implementation of our model is done with theano [21].

The answer is predicted according to whether the option is consistent with the document meaning for multiple-choice task, so we only evaluate our system performance in terms of precision (P = right_options/sum_options).

5.2. Results on MCTest Dataset. To verify the effectiveness of our proposed model, we test firstly our model on public datasets. Table 2 presents the performance of feature engineering and neural methods on the MCTest test set. The first four rows represent feature engineering methods and the last four rows are neural methods. As we can see the feature engineering methods outperform the neural methods significantly. One possible reason is that the neural methods suffered from the relative lack of training data. So we are going to analyze the related feature and add it to our neural network model in future work.

Table 3: Comparison of different reader model on three testing datasets.

Method	BCEETest	SBCEETest1	SBCEETest2
HABCNN-TE	0.428	0.442	0.438
Match Reader	0.461	0.452	0.455
SM Reader	0.495	0.491	0.499
CAS Reader	0.513	0.503	0.516
Sentence-Level Attention Reader (mode:max+avg)	0.581	0.535	0.578

For neural methods, the attentive reader [4] is implemented at word representation level and it is a deep model with thousands of parameters, so it performs poorly on MCTest. The neural reasoner [22] has multiple reasoning layers and all temporary reasoning affects the final answer representation. The HABCNN-TE [14] is convolutional architecture network. It can cut down on the parameter count, but the context representation can not be presented enough. Our method addresses the problems of the above methods. Firstly, the recurrent architecture network also cuts down on the parameter count and it can present the context representation at sentence level. Then, we use the max+avg method to reduce the impact of all snippets. Experimental results also demonstrate that our method performs better than the other three neural methods.

5.3. Results on Chinese Reading Comprehension Datasets. We have set four baselines for Chinese reading comprehension datasets. One is the HABCNN-TE method which is the most optimal method on MCTest datasets and the other three are as follows.

(i) The first baseline is inspired by Cui et al. [7]. We use the consensus attention-based neural network (called CAS Reader) for word of document and option. The model computes the attention of each document word directly, in respect to each option word at time t. The final consensus attention of option is computed by a merging function.

(ii) The second baseline uses a sliding window and matches a bag of words constructed from the document and the option, respectively (called Match Reader). This baseline is inspired from Zhang et al. [23].

(iii) The third baseline is the sentence similarity measure model (called SM Reader). The similarity is presented by the cosine similarity between the document sentence and the option sentence. The sentence representation is taken from Tai et al. [24]. The experimental results are given in Table 3.

The results on three test sets show that our sentence-level attention reader gives competitive results among various state-of-the-art baselines. We can observe that the accuracy in BCEETest outperforms the other test set. A possible reason can be that the college entrance examination is more standardized than that of the simulation. Also, we have noticed that the performance of the sentence-level model is better than the word-level model. For example, in BCEETest set, the SM Reader (sentence-level) outperforms the Match Reader (word-level) by 3.4% and The Sentence-Level Attention Reader (sentence-level) outperforms the CAS Reader (word-level) by 4.9% in precision, respectively.

Table 4: Results of different merging function.

	BCEETest	SBCEETest1	SBCEETest2
Mode:avg	0.562	0.513	0.550
Mode:sum	0.554	0.503	0.531
Mode:max	0.492	0.496	0.442
Mode:max+avg	0.581	0.535	0.578

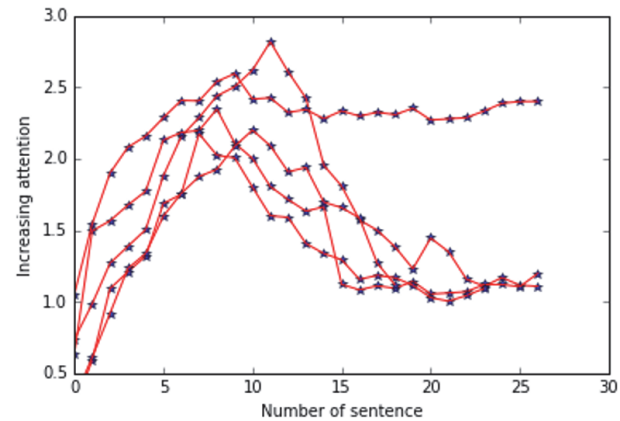


Figure 3: Experiment about the top N.

In experimenting we find out that the number of related sentences with the option is very important. So we also evaluate different merging functions as CAS Reader. The results are shown in Table 4. From the results, we can see that the avg and sum methods outperform the max method. A possible reason can be that the max method is equivalent to one sentence of document instead of the original document and a lot of information is lost. However, doing it achieves the best performance in which all sentences in document are used in the model. In order to measure it, we also use the max+avg method as the merging function. The "max" denotes the top N sentences and the "avg" denotes the average of top N sentences. In comparison with the avg method, the accuracy of the max+avg method increased by around 2% on three datasets. And this result is consistent with error analysis in Section 5.5. We suspect that some sentences interfere with the final answer as negative factor. Figure 3 shows the experiment about top N. We select randomly 5 options to do the experiment from the 13 Beijing college entrance examination papers (BCEETest). As we can see, the attention will not continue to increase in around 10. So N is set to 10 in our model. As shown in Box 2. The bold word

"Ruins" is a derogatory term that it is irrelevant to cultural and aesthetic in many Chinese mind, and interpretation of the word "ruins" is only a "city and village are changed into desolate places by destruction or natural disasters" in the "Modern Chinese Dictionary"; There is no fault for the interpretation, but it is not enough if it is measured by world knowledge. *In Europe, the meaning of "ruins" has been enriched and expanded since modern times. It has been endowed with the connotation of culture and aesthetics, and has become an academic concept.....*

C_i="One of the purposes of this paper is to correct the misunderstanding of the term "ruins" in the modern Chinese dictionary."

Box 2: Example of related sentences with the choice.

Table 5: Results of two sentence representation models.

	BCEETest	SBCEETest1	SBCEETest2
LSTM sentence model	0.518	0.495	0.489
Sentence composition model	0.483	0.506	0.522
Fusion model	0.581	0.535	0.578

denotes the most related sentences with the choice c_i; the italic word has a little relation with the choice c_i; the "....." is not relation.

5.4. Sentence Representation Model Analysis. In this paper, we use two models for the sentence representation, which are LSTM sentence model and sentence composition model [17]. Therefore, we have tested the contribution of the two models to the final model, respectively. The results are shown in Table 5.

The results on three test sets show that the precision of the fusion model is better than that of any single model. Therefore, we use the fusion model in sentence-level attention neural network.

5.5. Error Analysis. To better evaluate the proposed approach, we perform a qualitative analysis of its errors. Two major errors are revealed by our analysis, as discussed below.

(i) The positioning feature word (such as "The second paragraph…") often appears in the options. To further analyze the locating property of our model, we also examine the dependence of accuracy on the positioning feature word. And all sentences are replaced by related sentences of the positioning feature word in document. The accuracy has increased by about 3% on these three datasets. The positioning feature word we use is shown as follows.

[The end of paper; The second paragraph; The end paragraph; The end of paper; The first paragraph]

According to the above description, we will consider adding more features, such as location features, into our model in future work.

(ii) Our model may make mistakes when the option is expressed with emotion (such as "This paper not only contains historical interest, but also infiltrated the concern of reality and express the author's desire to enhance the cultural quality of the nation."). It is very difficult to calculate the attention between the option emotion and the document emotion. To handle such case correctly, our model will consider the emotion feature in future work. We have about more than 500 emotion feature words, like "thought provoking", "directly express one' s mind", and so forth.

6. Conclusion

In this paper, we introduce a sentence-level neural network model to handle the multiple-choice Chinese reading comprehension problems. The experimental results show that our model gives a state-of-the-art accuracy on all the evaluated datasets. We also use the max+avg method as the merging function and the accuracy of the max+avg method increased by about 2%. Furthermore, we analyze the positioning feature word and find that the accuracy increased by about 3%.

The future work will be carried out in the following aspects. First, we would like to extend our Chinese reading comprehension datasets and release it. Second, we are going to analyze the emotion feature and add it to our neural network model.

Data Availability

The Chinese reading comprehension data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (nos. 61772324, 61673248), the Shanxi Natural Science Foundation of China (no. 201601D102030), and Program for Fostering Talents of Shanxi Province Joint Postgraduate Training Base (nos. 2017JD05, 2018JD01).

References

[1] D. Chen, J. Bolton, and C. D. Manning, "A thorough examination of the CNN/daily mail reading comprehension task," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, pp. 2358–2367, August 2016.

[2] Y. Z. Liu, S. M. Sun, and L. Y. X. Kand Ruobing, "Knowledge representation learning: a review," *Journal of Computer Research and Development*, vol. 53, no. 2, pp. 247–261, 2016.

[3] R. Matthew, J. C. Christopher, and R. Erin, "MCTest: a challenge dataset for the open-domain machine comprehension of text," in *Proceedings of the 2013 on Empirical Methods in Natural Language Processing*, pp. 193–203, 2013.

[4] H. Felix, B. Antoine, C. Sumit, and W. Jason, "The goldilocks principle: reading children s books with explicit memory representations," 2015, https://arxiv.org/abs/1511.02301.

[5] R. Pranav, Z. Jian, L. Konstantin, and L. Percy, "SQuAD: 100,000+ Questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

[6] J. Mandar, C. Eunsol, SW. Daniel, and Z. Luke, "TriviaQA: a large scale distantly supervised challenge dataset for reading comprehension," in *Proceeding of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1601–1611, 2017.

[7] Y. M. Cui, T. Liu, and Z. P. Chen, "Consensus attention- based neural networks for Chinese reading comprehension," 2016, https://arxiv.org/help/prep.

[8] Y. Cui, Z. Chen, S. Wei, S. Wang, T. Liu, and G. Hu, "Attention-over-attention neural networks for reading comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 593–602, 2017.

[9] M. H. Karl, K. Tomas, G. Edward et al., "Teaching machines to read and comprehend," *In Advances in Neural Information Processing Systems*, pp. 1684–1692, 2015.

[10] J. M. Karl PRaymond, "Using sentence-level LSTM language models for script inference," in *Proceeding of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 279–289, 2016.

[11] S. Mrinmaya, D. Avinava, P. X. Eric, and R. Matthew, "Learning answer-entailing structures for machine comprehension," in *In Proceeding of the 53th Annual Meeting of the Association for Computational Linguistics*, pp. 239–249, 2015.

[12] H. Wang, B. Mohit, G. Kevin, and A. M. David, "Machine comprehension with syntax, frames, and semantics," in *Proceedings of the 53th Annual Meeting of the Association for Computational Linguistics*, pp. 700–706, 2015.

[13] T. Adam, Z. Ye, and Y. Xingdi, "A parallel-hierarchical model for machine comprehension on sparse," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 432–441, 2016.

[14] W. Yin, S. Ebert, and H. Schütze, "Attention-Based Convolutional Neural Network for Machine Comprehension," in *Proceedings of the Workshop on Human-Computer Question Answering*, pp. 15–21, San Diego, California, June 2016.

[15] K. Rudolf, S. Martin, B. Ondrej, and K. Jan, "Text understanding with the attention sum reader network," 2016, https://arxiv.org/abs/1603.01547.

[16] D. Bhuwan, H. Liu X, and L. Yang Z, "Gated-attention readers for text comprehension," in *Proceeding of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1832–1846, 2017.

[17] Y. L. Wang, "Sentence composition model for reading comprehension," *Journal of Computer Application*, vol. 37, no. 6, pp. 1741–1746, 2017.

[18] D. Z. Matthew, "Adadelta: an adaptive learning rate method," 2012, https://arxiv.org/abs/1212.5701.

[19] M. Tomas, C. Kai, C. Greg, and D. Jeffrey, "Efficient estimation of word representations in vector space," in *Proceedings of the In Proceedings of workshop at ICLR*, pp. 1–12, 2013.

[20] S. Nitish, E. H. Geoffrey, K. Alex, S. Ilya, and S. Ruslan, "Dropout, a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[21] Theano Development Team, "Theano: a python framework for fast computation of mathematical expressions," 2016, https://arxiv.org/abs/1605.02688.

[22] P. Baolin, L. Zhengdong, L. Hang, and W. Kanfai, "Toward neural network-based reasoning," 2015, https://arxiv.org/abs/1508.05508.

[23] Zhang Z. C., Z. Yu, and T. Liu, "Answer sentence extraction of reading comprehension based on shallow sematic tree kernel," *Journal of Chinese Information Processing*, vol. 22, no. 1, pp. 80–86, 2008.

[24] T. Kaisheng, S. Richard, and D. M. Christopher, "Improved semantic representations from tree-structured long short-term memory networks," in *Proceedings of the 53th Annual Meeting of the Association for Computational Linguistics*, pp. 1556–1566, 2015.

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 1827074, 21 pages
https://doi.org/10.1155/2018/1827074

Review Article

A Bibliometric Review of Natural Language Processing Empowered Mobile Computing

Xieling Chen,[1] **Ruoyao Ding**,[2] **Kai Xu**,[3] **Shan Wang**,[4] **Tianyong Hao**,[5] **and Yi Zhou**[6]

[1]*College of Economics, Jinan University, Guangzhou, China*
[2]*School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, China*
[3]*School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China*
[4]*Department of Chinese Language and Literature, University of Macau, Macau SAR, China*
[5]*School of Computer, South China Normal University, Guangzhou, China*
[6]*Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China*

Correspondence should be addressed to Tianyong Hao; haoty@gdufs.edu.cn and Yi Zhou; zhouyi@mail.sysu.edu.cn

Received 23 January 2018; Accepted 5 April 2018; Published 28 June 2018

Academic Editor: Javier Prieto

Natural Language Processing (NLP) empowered mobile computing is the use of NLP techniques in the context of mobile environment. Research in this field has drawn much attention given the continually increasing number of publications in the last five years. This study presents the status and development trend of the research field through an objective, systematic, and comprehensive review of relevant publications available from Web of Science. Analysis techniques including a descriptive statistics method, a geographic visualization method, a social network analysis method, a latent dirichlet allocation method, and an affinity propagation clustering method are used. We quantitatively analyze the publications in terms of statistical characteristics, geographical distribution, cooperation relationship, and topic discovery and distribution. This systematic analysis of the field illustrates the publications evolution over time and identifies current research interests and potential directions for future research. Our work can potentially assist researchers in keeping abreast of the research status. It can also help monitoring new scientific and technological development in the research field.

1. Introduction

With the development of mobile devices as well as the advances in wireless communication technologies, mobile computing is becoming a significantly important paradigm in today's world of networked computing systems [1]. Mobile computing enables a computer to be used normally while in the state of movement. Based on perceived situational information in personal and ubiquitous environments, mobile computing provides services automatically. With the rapid growth in use of mobile devices, far-reaching and diverse information is being produced rapidly and distributed instantly in digitized format [2]. A large amount of valuable information existing in unstructured texts are of great need of processing, such as web pages, short messages, Twitter/WeChat messages, etc. Natural Language Processing (NLP) focuses on the interactions between computers and natural language texts. NLP is capable of providing a computer program with the ability to process and understand unstructured texts. By automatically analyzing the meaning of user content to take appropriate actions, NLP can make applications smarter in the mobile environment.

NLP empowered mobile computing research field has attracted more and more interests from scientific community, witnessing 12 publications in 2000 to 55 publications in 2016 from Web of Science (WoS). Some representative examples are as follows. Chen et al. [3] applied the technique of multitask learning using deep neural networks to Mandarin-English code-mixing recognition. Three schemes of the auxiliary tasks were proposed to introduce the language

information to networks and to improve the prediction of language switching for the primary task of senone classification. The proposed schemes enhanced the recognition on both languages and reduced the relative overall error rates by 4.4% on average when dealing with real-world Mandarin-English corpus in mobile voice search. Ilayaraja et al. [4] presented a weighted association rule mining prefetching technique to determine the secondary service item, with the consideration of access frequency of services, semantic distance among the successive query request, and spatial distance between service instances and user context (e.g., position, service type, and query request time). Wong et al. [5] analyzed the students' vocabulary usage using a corpus analysis tool to identify and unpack the contextual conditions in which a mobile- and cloud-assisted Chinese language learning environment promoted key learning outcomes. Räsänen and Saarinen [6] proposed a method based on sparse hyperdimensional coding of sequence structures for sequence prediction. Their experiments suggested that the method was capable of capturing the relevant variable-order structure from the sequences. A NLP based tool MOTTE was developed by Puppala et al. [7] for extracting and structuring data in pathology reports automatically to support clinical solution applications. With an aim of screening information on human immunodeficiency virus/acquired immune deficiency syndrome, Adesina et al. [8] designed a monolingual short message services based system for the retrieval of frequently asked questions.

Bibliometric analysis is defined as the use of statistical methods on evaluating scholarly publications from an objective and quantitative perspective within a certain field [9]. Benefits of bibliometric analysis include (1) organizing information in a specific thematic field [10], (2) evaluating scientific developments in knowledge of a specific subject and assessing the scientific quality [11], (3) determining the impact of research funding, (4) comparing research performance across different affiliations and document changes in the research workforce, and (5) identifying emerging areas of research focus and predicting future research success [12]. As for researchers, especially newcomers, bibliometric analysis can assist them in (1) better selecting potential research topics, (2) demonstrating the values and impacts of their relevant works, (3) recognizing appropriate academic researchers to seek research collaboration, and (4) keeping abreast of new research status and new technological changes [13].

Bibliometric analysis has been widely applied to various fields for the measurement of quality and productivity of academic output and has demonstrated excellent effectiveness from long-term practice. Relevant researches mainly focused on revealing publication statistical characteristics, exploring the collaboration relationship, and uncovering research themes and their evolution. Some examples are as follows. Geng et al. [14] conducted a bibliometric survey of the research field of residential energy and greenhouse gas emissions for the purpose of uncovering research status. In their work, citation analysis was used to assess the influence of journals, countries, and authors, while network analysis was performed to evaluate the relationships among countries, authors, and keywords. Based on 117,340 obesity-related research publications indexed in Scopus database published from 1993–2012, Khan et al. [15] reported research trends and collaboration patterns in the field. Roig-Tierno et al. [16] conducted a bibliometric analysis on research publications with the application of qualitative comparative analysis (QCA). Their study revealed the differences in quantitative terms of the three variants of QCA. Albort-Morant and Ribeiro-Soriano [17] focused on the research development of business incubators. They sorted 445 publications from WoS according to bibliographic indicators such as research area and year of publication. Their study revealed the lack of publications on business incubators and highlighted the fragmented nature of research themes. Merigó and Yang [18] aimed at identifying relevant researches and the newest trends in field of operation research and management science. The analysis involved some influential journals, two hundred most cited publications, and productive and influential authors. Zhang et al. [19] quantitatively and qualitatively evaluated carbon tax related literature from 1989 to 2014 using bibliometric analysis. Their study demonstrated that the USA was the leading country and *the Vrije University Amsterdam* and *Massachusetts Institute of Technology* and *Stanford University* were the most productive affiliations in the research field. Randhawa et al. [20] conducted a systematic review of publications on open innovation (OI) research area using bibliometrics, cocitation analysis, and text mining. Three distinct areas within OI research were identified, i.e., firm-centric aspects of OI, management of OI networks, and role of users and communities in OI. In order to discover the worldwide trends in the research field of drying brick/tile, Yataganbaba and Kurtbaş [21] analyzed relevant patents in terms of, e.g., publication number, authorship and ownership, and international collaboration patterns. Merigó et al. [10] explored the research development trends in fuzzy sciences. Similar works have also been conducted in other fields, e.g., natural language processing [22], neuroimaging [23], and diabetes [24].

To the best of our knowledge, there is no scientific review of NLP empowered mobile computing research field currently. Thus, in this study, we conduct a bibliometric analysis on publications retrieved from WoS during the years 2000–2016 to explore the research status of the research field. The main objective is to address the following issues: (1) investigating publication statistical characteristics and publication collaborations, (2) exploring publication geographical distributions, (3) visualizing scientific collaboration relationships, and (4) reveling current hot research topic themes and research topic changes.

The rest of the paper is organized as follows. Section 2 introduces methods and materials. Bibliometric analysis results on retrieved research publications are reported in Section 3. Findings and discussion are shown in Section 4 while Section 5 summarizes the work.

2. Methods and Materials

Five different methods are applied to analyze research publications in the NLP empowered mobile computing field

retrieved from WoS. The details of the methods are described in Section 2.1 and the publication data is introduced in Section 2.2.

2.1. Methods

2.1.1. Descriptive Statistics Method. Descriptive statistics are brief descriptive coefficients that summarize a collection of information, which can be either a representation of the entire population or a sample. Descriptive statistics are commonly used as measures of central tendency and measures of variability. Measures of central tendency usually include mean, median, and mode, while measures of variability generally contain standard deviation, minimum and maximum variables, kurtosis, and skewness. These two measures use graphs, tables, and general discussions to simply describe data. This simplifies large amounts of data in a sensible way by presenting quantitative descriptions in a manageable form to help users understand the meaning of the data being analyzed.

In this study, descriptive statistics method was applied to acquire characteristics of the retrieved publications, including publication distribution by year, most influential publications, productive journals, authors, affiliations, and countries/regions, as well as co-authors, coaffiliation, and cocountry/region publication distribution and topic distribution by year.

2.1.2. Geographic Visualization Method. Geographic visualization or Geovisualization is a set of tools and techniques supporting the analysis of geospatial or spatial data, emphasizing knowledge construction over knowledge storage or information transmission. By combining technologies, e.g., image processing, simulation, and virtual reality, computers can help present information in a way that patterns can be found. Geovisualization can be applied to all the stages of problem-solving in geographical analysis, from development of initial hypotheses to knowledge discovery, analysis, presentation, and evaluation. According to Tobler's First Law of Geography [25], everything is related to everything else, but near things are more related than distant things. Through Geovisualization, we can use location as the key index variable and get related information which is previously unfound. Locations or extents in the earth space–time may be recorded as dates/times of occurrence. Longitude, latitude, and elevation are represented as X, Y, and Z coordinates, respectively.

In this study, we applied geographic visualization analysis to explore geographical distributions of publications in country/region level.

2.1.3. Social Network Analysis Method. Social network analysis is a process of investigating social structures using networks and graph theory [26]. It focuses on relationship structures, ranging from casual acquaintance to close bonds. Network structures are characterized in terms of nodes (items, individuals, or things within the network) with the edges or links (relationships or interactions) connecting the nodes. Researches using social network analysis have been undertaken in different areas, e.g., collaboration graphs [27], social media networks [28], and disease transmission [29]. These networks are often visualized through sociograms in which nodes are represented as points and edges are represented as lines. The social network analysis can help identify the individuals, teams, and units who play central roles, leverage peer support, and strengthen the efficiency and effectiveness of existing channels [30].

In this study, we applied social network analysis to explore the cooperation relationships for specific countries/regions, affiliations, and authors in the NLP empowered mobile computing research field. The cooperation among countries/regions, affiliations, and authors was visualized using interactive force directed networks. In the networks, nodes represented specific countries/regions, affiliations or authors, and lines indicated cooperation. The size of nodes represented publication numbers of a specific country, affiliation, or author. The width of lines reflected cooperation frequencies between two countries/regions, affiliations, or authors. The color indicated specific continent of a country/region, or specific country/region of an affiliation or author. Users could explore the cooperation relationships for specific countries/regions, affiliations, or authors by dynamically dragging the nodes.

2.1.4. Latent Dirichlet Allocation Method. Latent Dirichlet allocation (LDA), proposed by Blei [31], is a generative probabilistic model. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words, and topics are assumed to be uncorrelated.

LDA formally defines the following terms:

(1) A *word* is defined as an item from a vocabulary indexed by $\{1, \ldots, V\}$.

(2) A *document* is a sequence of N words denoted by $d = (w_1, \ldots, w_N)$.

(3) A *corpus* is a collection of M documents denoted by $D = \{d_1, \ldots, d_M\}$.

LDA assumes the following generation process:

(1) The term distribution β which contains the probability of a word occurring in a given topic is determined by $\beta \sim \text{Dirichlet}(\delta)$.

(2) The proportions θ of the topic distribution for a document d are determined by $\theta \sim \text{Dirichlet}(\alpha)$.

(3) For each word w_i in the document d, a topic is chosen by the distribution $z_i \sim \text{Multinomial}(\theta)$ and a word is chosen from a multinomial probability distribution conditioned on the topic $z_i : p(w_i \mid z_i, \beta)$.

As for variational expectation-maximization (VEM) estimation, the log-likelihood for one document $d \in D$ is given by

$$\ell(\alpha, \beta) = \log(p(d \mid \alpha, \beta))$$

$$= \log \int \left\{ \sum_{z} \left[\prod_{i=1}^{N} p\left(w_i \mid z_i, \beta\right) p\left(z_i \mid \theta\right) \right] \right\} \cdot p\left(\theta \mid \alpha\right) d\theta \tag{1}$$

Gibbs sampling defines a Markov chain in the space of possible variable assignments such that the stationary distribution of the Markov chain is the joint distribution over variables. Thus, it is a Markov Chain Monte Carlo method [32]. Its aim is to construct a Markov chain converging to the target probability distribution in the high dimensional model and then the sample distribution closest to the target probability distribution will be extracted. The log-likelihood for Gibbs sampling can be obtained through

$$\log\left(p\left(d \mid z\right)\right) = k \log\left(\frac{\Gamma\left(V\delta\right)}{\Gamma\left(\delta\right)^{V}}\right) + \sum_{K=1}^{k} \left\{ \left[\sum_{j=1}^{V} \log\left(\Gamma\left(n_K^{(j)} + \delta\right)\right) \right] - \log\left(\Gamma\left(n_K^{(\cdot)} + V\delta\right)\right) \right\}. \tag{2}$$

The perplexity, as shown in (3), is often used to evaluate the models on held-out data and is equivalent to the geometric mean per-word likelihood. The less the perplexity is, the better the model is.

$$\text{perplexity}\left(d\right) = \exp\left\{ -\frac{\log\left(p\left(d\right)\right)}{\sum_{d=1}^{D} \sum_{j=1}^{V} n^{(jd)}} \right\}. \tag{3}$$

In (4), $n^{(jd)}$ denotes how often the jth term occurs in the dth document. If the model is fitted through Gibbs sampling, the likelihood can be determined for the perplexity using

$$\log\left(p\left(d\right)\right) = \sum_{d=1}^{D} \sum_{j=1}^{V} n^{(jd)} \log\left[\sum_{K=1}^{k} \theta_K^{(d)} \beta_K^{(j)} \right] \tag{4}$$

Additionally, estimation using Gibbs sampling requires specification of values for the parameters of the prior distributions.

In this study, topic discovery and distribution were analyzed using LDA models with the following steps:

(1) We assigned the weights of segmented author keywords and Keywords Plus, publication title, and abstract as 0.4, 0.4 and 0.2, respectively, as determined in our former experiment [13].

(2) Term Frequency-Inverse Document Frequencies (TF-IDF) were used to filter out unimportant terms. As one of the most popular term-weighting schemes, TF-IDF increases proportionally to the number of times a term appears in a publication but is often offset by the frequency of the term in the whole collection of publications. We calculated the TF-IDF values of all terms to sort the terms. By manually examining these ranked terms, we defined a threshold as 0.1 empirically. Only the terms with a TF-IDF value greater than the threshold were kept for further analysis.

(3) Through sampling, 16 different topic numbers were set to $c(2:10, 15, 20, 40, 50, 80, 150, 250)$. For each topic number, 10-fold cross-validation was used to evaluate model performance. Specifically, dataset was split into 10 test datasets to conduct multiple runs. Perplexity criteria were used to select optimal topic number. α for Gibbs sampling was initialized as the mean value of α values for model fitting using VEM with the optimal topic number.

(4) With an initialized α and the optimal topic number, we adopted Gibbs sampling and VEM method to estimate the LDA model.

(5) By matching the topics detected by VEM and Gibbs sampling based on Hellinger distance, the best matches with the smallest distance could be identified. Hellinger distance is calculated as (5), in which P and Q denote two probability measures.

$$H^2\left(P, Q\right) = \frac{1}{2} \int \left(\sqrt{dP} - \sqrt{dQ}\right)^2. \tag{5}$$

2.1.5. Affinity Propagation Clustering Method. Affinity Propagation (AP) algorithm was proposed by Frey and Dueck [33]. It is a technique for data clustering based on message passing. AP does not require the predefined number of clusters. It identifies cluster centers, or exemplars as representative members of clusters. Initially, all nodes are considered as exemplars. "Preference" is used to reflect how likely one node is chosen as an exemplar. If no prior knowledge is available, all nodes will be assigned the same preference value. AP has been shown to be more efficient and effective in cluster identification than traditional clustering methods, e.g., k-means [34].

AP algorithm takes $s(i, j)$ as function of similarity to reflect the fitness of the data point j being the exemplar of data point i. The aim of AP is to maximize the similarity $s(i, j)$ between every data point i and its chosen exemplar j. Each node i also has a self-similarity $s(i, i)$. Individual data points initialized with a larger self-similarity are more likely to become exemplars. All data points are equally likely to be exemplars when they are initialized with the same constant self-similarity. The number of clusters produced will be increased and decreased accordingly with this common self-similarity input.

There are two types of messages contained in this technique. The responsibility $r(i, j)$ is directed from i to candidate exemplar j. It indicates how well suited j is to be i's exemplar, taking into consideration competing potential exemplars. The availability $a(i, j)$ is sent from candidate exemplar j back to i. It indicates j's desire to be an exemplar for i based on supporting feedback from other data points. Both the self-responsibility $r(i, i)$ and the self-availability $a(i, i)$ can

Table 1: The query used to retrieve research publications in the NLP empowered mobile computing field from WoS.

TS=((“natural language processing” OR “NLP” OR “semantic analysis” OR “bag of words” OR “word sense disambiguation” OR “named entity recognition” OR “NER” OR “sentiment analysis” OR “information extraction” “tokenization” OR “stemming” OR “lemmatization” OR “corpus” OR “stop words” OR “parts-of-speech” OR “language modeling” OR “n-grams” OR “syntactic analysis” OR “information retrieval” OR “language model”) AND (“mobile computing” OR “mobile” OR “smart device” OR “smartphone” OR “cellphone” OR “telephony device” OR “Cellular network” OR “Android” OR “iOS” OR “phone”))

reflect accumulated evidence that i is an exemplar. The update formulas for responsibility and availability are as follows:

$$r(i,j) \longleftarrow s(i,j) - \max_{j' \text{ s.t. } j' \neq j} \left\{ a\left(i,j'\right) + s\left(i,j'\right) \right\}$$

$$a(i,j) \underset{i \neq j}{\longleftarrow} \min \left\{ 0, r(i,j) + \sum_{\forall i' \notin \{i,j\}} \max \left\{ 0, r(i',j) \right\} \right\} \quad (6)$$

$$a(j,j) \longleftarrow \sum_{i' \text{ s.t. } i' \neq j} \max \left\{ 0, r\left(i',j\right) \right\}.$$

Responsibility and availability of message updates are $m_{\text{new}} = \lambda m_{\text{old}} + (1-\lambda) m_{\text{new}}$, where λ is a weighting factor between 0 and 1. In AP, the clustering is complete when the messages converge. Also, AP algorithm is able to determine when a specific data point has converged to cluster head status in its given cluster. A point becomes the cluster head when its self-responsibility plus self-availability becomes positive. Upon convergence, each node i's cluster head can be calculated using

$$CH_i = \arg\max_j \left\{ a(i,j) + r(i,j) \right\}. \quad (7)$$

In our study, with the basis of term-topic posterior probability matrix, we applied AP clustering method for the cluster analysis of the topics identified by the LDA method.

2.2. Materials. Web of Science, as the most authoritative citation database, was used as the data source for retrieving research publications in the NLP empowered mobile computing field. First of all, a list of keywords related to the “natural language processing” and “mobile computing” was determined by a domain expert. With “Science Citation Index Expanded” and “Social Sciences Citation Index” as indexes, publications used in this study were identified using the specific query in Table 1. 716 publications in “article” type during years 2000–2016 were obtained. Citations counted to September 8th, 2017 were considered for each publication.

The raw data of the 716 publications were downloaded as plain text. Key elements including title, author, journal, publication date, subject category, language, funding, author keywords, Keywords Plus, abstract, and author address, as well as number of citations, pages, and references, were extracted. In order to ensure they were closely related to the research field, manual verification was conducted by a domain expert on each publication. 471 publications were identified as relevant for analysis eventually. Further, corresponding affiliations and countries/regions were identified out from author address information. Key terms were extracted from author keywords, Keywords Plus, title, and abstract.

The statistical characteristics of the publications are shown as Table 2. The average page number of the publications is 15.66 and the average reference number of the publications is 33.29. There are 48 subject categories included, where the top 3 categories are computer science (38.76%), engineering (16.27%), and telecommunications (10.98%).

The distribution characteristics of the 471 publications are shown in Figure 1. Figure 1(a) shows the distributions of the numbers of countries/regions, affiliations, authors, and funds. Figure 1(b) shows the distributions of the numbers of keywords, pages, and references. The distribution of the number of title characters is shown in Figure 1(c). In Figure 1(d) the right bottom illustrates the distribution of the number of abstract characters.

3. Results

3.1. Publication with Year. The total publications, total citations, average number of citations per publication, and the number of annual citations are demonstrated in Figure 2. The results show that the research in the NLP empowered mobile computing field exhibits an overall upward trend in fluctuation (from 12 publications in 2000 to 55 publications in 2016). The publication number presents a stable increasing trend since 2010. Based on the data for years 2010–2016, we developed a regression model by setting the independent variables as *time/1000* and *(time/1000)*^{2}. The estimated regression model is calculated as $y = 6.7143 * 10^3 - 1.34777 * 10^4 x$. The adjusted goodness-of-fit $\overline{R^2}$ of the model is 0.9468. With the regression model, publication number in 2017 is predicted as 65, while the actual number of publications on WoS in 2017 is 66. The trend of citations does not keep step with publication number, and extreme values appear in 2002 as 431, 2007 as 503, and 2010 as 490. The average number of citations per publication is calculated as *total citations/total publications*. It shows an overall downward trend in fluctuation from 21.92 in 2000 to 2.53 in 2016. We eliminated the influence of duration since first publication using the formula: *the number of annual citations* (*C/Y*) = *total citations*/(*2016 + 1-publishing year*). The number of annual citations increases in fluctuation from 15.47 in 2000 to 139 in 2016.

3.2. Productive Journals. The top 11 contributing journals in the research field are presented in Table 3. These journals contribute about 21% of the total publications and 29.20% of the total citations. The most productive 3 are *IEEE/ACM Transactions on Audio Speech and Language Processing* (25

Table 2: The statistical characteristics of the 471 publications.

Characteristics	Statistics
Total #pub.	471
#pub. with author keywords or Keywords Plus	412
#unique publication sources	287
#unique countries (or regions)/first countries (or regions)	60; 52
#unique affiliations/first affiliations	544; 345
#unique authors/first authors/last authors	1,408; 451; 441
Average #citations	10.42
Average #countries (or regions) in one pub.	1.25
Average #affiliations in one pub.	1.64
Average #authors in one pub.	3.27
Average #funds in one pub.	0.73
Average #pages in one pub.	15.66
Average #references in one pub.	33.29
Average #author keywords or Keywords Plus	6.81
Average #words/characters in title	10.57; 80.13
Average #words/characters in abstract	186.40; 1,265.58
Language distribution	English (98.73%); Estonian (0.42%); French (0.42%); Spanish (0.21%); Afrikaans (0.21%)
Subject category distribution (Top 10)	Computer Science (38.76%); Engineering (16.27%); Telecommunications (10.98%); Acoustics (5.82%); Information Science & Library Science (2.78%); Linguistics (2.51%); Psychology (2.12%); Operations Research & Management Science (1.85%); Business & Economics (1.32%); Communication (1.32%)
Top 10 terms in author keywords and Keywords Plus	Mobile (30.36%); Information (22.08%); Retrieval (16.77%); Recognition (16.56%); System (14.86%); Speech (14.01%); Model (12.10%); Network (12.10%); Language (11.04%); Analysis (9.98%)
Top 10 terms in titles	Mobile (34.18%); Information (17.83%); System (12.53%); Retrieval (12.10%); Recognition (10.62%); Speech (8.70%); Network (8.28%); Model (7.86%); Language (7.22%); Environment (6.37%)
Top 10 terms in abstracts	Mobile (66.67%); Information (56.90%); Paper (55.41%); System (48.20%); Result (46.07%); Data (38.00%); User (38.00%); Model (37.37%); Device (32.70%); Retrieval (31.42%)

Table 3: Top 11 contributing journals in the NLP empowered mobile computing research field.

Rank	Journals	SC	TP	% *P*	TC	ACP	*H*	≥10	T100
1	IEEE/ACM Transactions on Audio Speech and Language Processing	A; E	25	5.31	447	17.88	11	12	11
2	Speech Communication	A; CS	11	2.34	179	16.27	6	6	5
3	Computer Speech and Language	CS	10	2.12	93	9.30	6	5	3
4	Expert Systems with Applications	CS; E; OR&MS	8	1.70	320	40.00	8	7	5
4	IEEE Transactions on Consumer Electronics	E; T	8	1.70	44	5.50	5	1	0
6	Mobile Information Systems	CS; T	7	1.49	95	13.57	3	2	2
6	Multimedia Tools and Applications	CS; E	7	1.49	71	10.14	3	1	1
6	Personal and Ubiquitous Computing	CS; T	7	1.49	67	9.57	4	3	1
9	Information Sciences	CS	6	1.27	85	14.17	5	3	3
10	EURASIP Journal on Wireless Communications and Networking	E; T	5	1.06	22	4.40	2	1	1
10	IEICE Transactions on Information and Systems	CS	5	1.06	11	2.20	2	0	0

Notice. Journal *IEEE Transactions on Audio Speech and Language Processing changed name as IEEE/ACM Transactions on Audio, Speech, and Language Processing* in 2013, and journal *IEEE Transactions on Speech and Audio Processing* ceased publication in 2005, and the current retitled publication is *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. Therefore, publications from these 2 journals were combined as published by *IEEE/ACM Transactions on Audio, Speech, and Language Processing*; *Abbreviations*. SC: subject categories only with NLP empowered mobile computing research (A: acoustics; E: engineering; CS: computer science; OR&MS: operations research and management science; T: telecommunications); TP: total publications; % *P*: percentage of the publications; TC: total citations; ACP: average number of citations per publication, calculated as TC/TP; *H*: *H*-index; ≥10: number of publications with citations ≥10; T100: number of publications in the top 100 most influential publications.

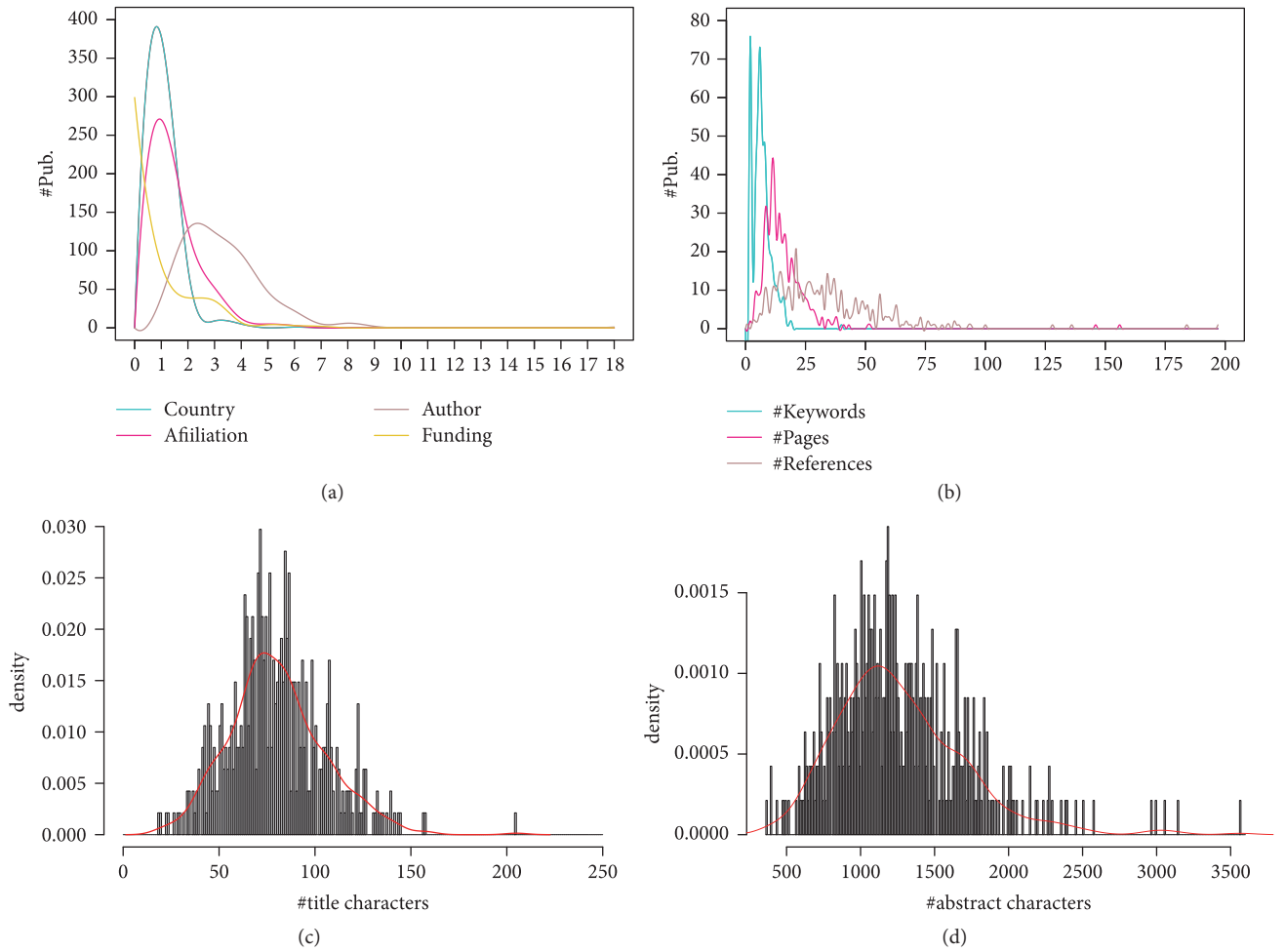


Figure 1: Distribution characteristics of the 471 publications.

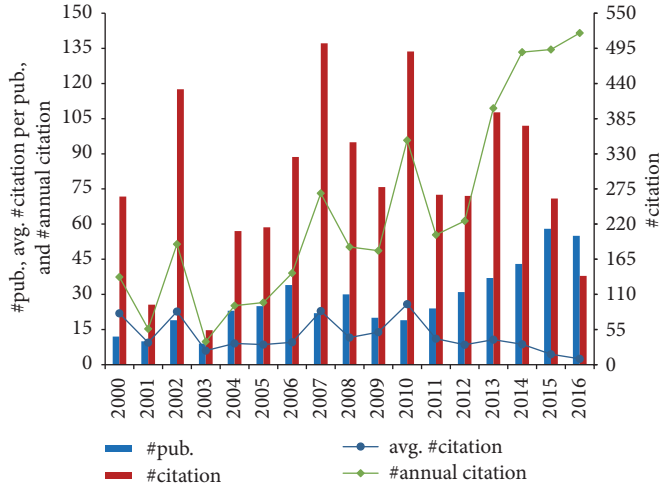
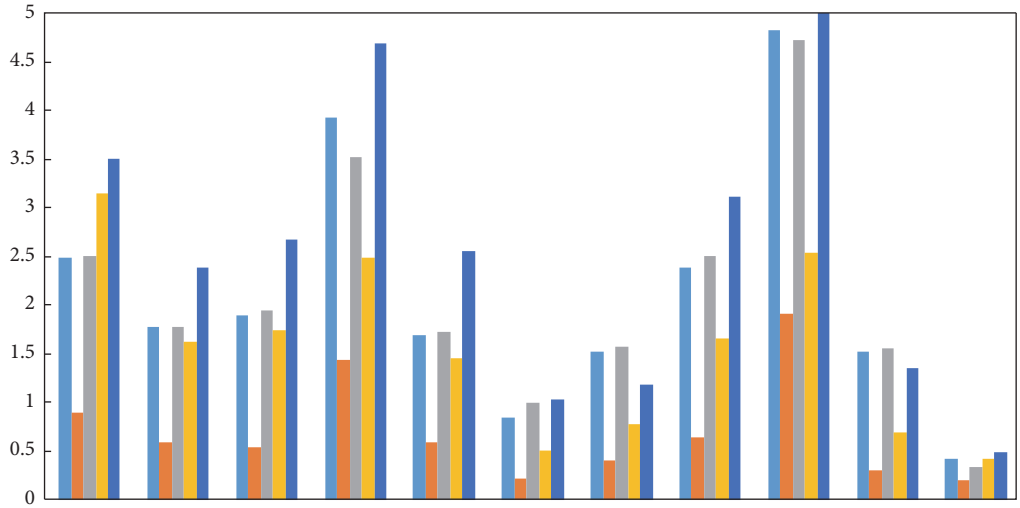


Figure 2: The statistics of the 417 publications (the light blue bars indicate total publications and the red bars indicate total citations. The dark blue line indicates average citations per publication and the green line indicates annual citations).



IF	2.491	1.768	1.9	3.928	1.694	0.849	1.53	2.395	4.832	1.529	0.411
SJR	0.813	0.593	0.536	1.433	0.587	0.219	0.396	0.648	1.91	0.304	0.191
5-Year IF	2.501	1.769	1.938	3.526	1.725	1	1.572	2.512	4.731	1.558	0.332
SNIP	3.143	1.628	1.75	2.492	1.447	0.509	0.775	1.651	2.537	0.698	0.419
CiteScore	3.5	2.38	2.67	4.7	2.56	1.03	1.19	3.12	5.37	1.36	0.49

Figure 3: Comparisons of IF, SJR, 5-Year IF, SNIP, and CiteScore for the top 11 productive journals for year 2016.

publications, 447 citations, 17.88 ACP, and 11 H-index), *Speech Communication* (11 publications, 179 citations, 16.27ACP, 6 H-index), and *Computer Speech and Language* (10 publications, 93 citations, 9.30 ACP, 6 H-index). *Expert Systems with Applications* has the highest ACP of 40.00. We found that 32 of the 100 most influential publications are published in the 11 journals. According to subject category of these 11 journals, *computer science* possesses the widest influence in the research field.

In order to better measure the overall scientific importance of these 11 journals, 5 assessment indicators acquired from Scientific Journal Rankings were used, including Impact Factor (IF), SCImago Journal Rank (SJR), 5-Year IF, Source Normalized Impact per Paper (SNIP), and CiteScore. IF is a measure for reflecting the yearly average number of citations to recent publications published in a journal. It is the primary and widely used indicator on assessing one journal's significance. SJR is a measure of scientific influence of scholarly journals. It accounts for both the number of citations received by a journal and the importance or prestige of the journals where such citations come from. 5-Year IF is calculated by dividing the number of citations to the journal in a given year by the number of publications published in that journal in the previous five years. SNIP is defined as the ratio of the journal's citation count per publication and the citation potential in its subject field. CiteScore index, launched by Elsevier in December 2016, is calculated as the ratio of total citations received in a given year by all publications published in a given journal in three previous years and the number of publications published in the journal in three previous years.

Therefore, the 11 productive journals were compared by using their IF, SJR, 5-Year IF, SNIP, and CiteScore for year 2016, as shown in Figure 3. As for IF, SJR, and CiteScore, the top 3 are *Information Sciences* (IF 4.832, SJR 1.91, and CiteScore 5.37), *Expert Systems with Applications* (IF 3.928, SJR 1.433, and CiteScore 4.7), and *IEEE/ACM Transactions on Audio Speech and Language Processing* (IF 2.491, SJR 0.813, and CiteScore 3.5). As for 5-Year IF, the top 3 are *Information Sciences* (5-Year IF 4.731), *Expert Systems with Applications* (5-Year IF 3.526), and *Personal and Ubiquitous Computing* (5-Year IF 2.512). As for SNIP score, the top 3 are *IEEE/ACM Transactions on Audio Speech and Language Processing* (SNIP 3.143), *Information Sciences* (SNIP 2.537), and *Expert Systems with Applications* (SNIP 2.492).

3.3. Most Influential Publications. The number of citations reflects the popularity and influence of a publication in the scientific community [10]. Thus, we used the total citations as a measurement of influence. There are 69 and 129 publications with the number of citations ≥20 and ≥10. Top 15 most influential publications are listed in Table 4. The publication by Miao et al. [35] in 2010 (376 citations) is the most influential one, followed by [36] published by MacKenzie and Soukoreff in 2002 (172 citations) and [37] by Strayer and Drews in 2007 (148 citations). We further consider the number of annual citations of the 15 publications. The top 3 publications measured by this indicator are [38] by Cao et al.

Table 4: Top 15 most influential publications in the NLP empowered mobile computing research field.

Rank	Title	Author/s	Year	TC	C/Y
1	Energy-Efficient Link Adaptation in Frequency-Selective Channels	Miao G. W., et al.	2010	376	53.71
2	Text Entry for Mobile Computing: Models and Methods, Theory and Practice	MacKenzie I. S.; Soukoreff R. W.	2002	172	11.47
3	Cell-Phone-Induced Driver Distraction	Strayer D. L.; Drews F. A.	2007	148	14.80
4	A Vector Space Modeling Approach to Spoken Language Identification	Li H. Z., et al.	2007	116	11.60
5	Context-Aware System for Proactive Personalized Service Based on Context History	Hong J. Y., et al.	2009	91	11.38
6	More than Words: Social Networks' Text Mining for Consumer Brand Sentiments	Mostafa M. M.	2013	88	22.00
7	The Effect of Mobility-Induced Location Errors on Geographic Routing in Mobile Ad Hoc and Sensor Networks: Analysis and Improvement Using Mobility Prediction	Son, D. J., et al.	2004	77	5.92
8	A Personalized Tourist Trip Design Algorithm for Mobile Tourist Guides	Souffriau W., et al.	2008	76	8.44
9	D'Agents: Applications and Performance of a Mobile-Agent System	Gray R. S., et al.	2002	73	4.87
10	Optical Encryption and QR Codes: Secure and Noise-Free Information Retrieval	Barrera J. F., et al.	2013	64	16.00
11	Text-Dependent Speaker Verification: Classifiers, Databases and RSR2015	Larcher A., et al.	2014	60	20.00
12	A Location-Aware Recommender System for Mobile Shopping Environments	Yang W. S., et al.	2008	59	6.56
12	An Application of Reinforcement Learning to Dialogue Strategy Selection in a Spoken Dialogue System for Email	Walker M. A.	2000	59	3.47
14	Landmark Recognition with Compact BoW Histogram and Ensemble ELM	Cao J. W., et al.	2016	56	56.00
14	Mobile-Agent Coordination Models for Internet Applications	Cabri G., et al.	2000	56	3.29

Abbreviations. TC: total number of citations during 2000 and 2016; C/Y: the number of annual citations.

published in 2015 ($C/Y = 56$), [35] by Miao et al. in 2010 ($C/Y = 53.71$), and [39] by Mostafa in 2013 ($C/Y = 22$). These 3 publications rank 14th, 1st, and 6th, respectively, according to total citations.

3.4. Productive Authors and Affiliations. From the 471 publications, there are 1,408 authors. 451 of them are first authors and 441 are last authors. 20 authors have 3 or more publications, and 98 authors have 2 or more publications. 20 most productive authors are listed in Table 5. According to the result, the most productive authors are *Chen, Tao* from Singapore (4 publications supported by 4 funds, 108 citations, 27 ACP, and 4 H-index) and *Mizzaro, Stefano* from Italy (4 publications, 45 citations, 11.25 ACP, and 3 H-index). *Chen, Tao* is listed as first author of 3 publications and all the 3 publications appear in top 100 most influential publications. *Mizzaro, Stefano* cooperates with others in all his 4 publications and 1 publication appears in the top 100. As for the ranking based on citation number, the top 3 productive authors are *Lee, Chin-Hui* from the USA (173 citations and 57.67 ACP), *Chen, Tao* from Singapore (108 citations and 27 ACP), and *Xie, Xing* from China (51 citations and 17 ACP). Ranking based on the ACP indicator yields the same result. *Kim, Harksoo* from South Korea achieves the most funding supports, i.e., 7 for his 3 publications.

544 affiliations from 60 countries/regions have publications in the NLP empowered mobile computing research field. Table 6 lists 15 most productive affiliations. Among them, 5 are from the USA, 3 from China, 2 from Taiwan, 1 from India, 1 from Italy, 1 from South Korea, 1 from Singapore, and 1 from England. The top 4 most productive affiliations a *Nanyang Technological University* from Singapore (8 publications, 87 citations, 10.88 ACP, and 5 H-index), *Tsinghua University* from China (8 publications, 42 citations, 5.25 ACP, and 4 H-index), *Microsoft Research Asia* from China (7 publications, 115 citations, 16.43 ACP, and 5 H-index), and *National Taiwan University* from Taiwan (7 publications, 83 citations, 11.86 ACP, and 5 H-index). *Nanyang Technological University* cooperates with others in 5 publications and serves as first affiliation in 4 of them. 3 of these 5 publications appear in the list of top 100 most influential publications. *Tsinghua University* cooperates with others in 4 publications and serves as first affiliation in all 8 publications. These 8 publications are supported by 21 funds. As for the ranking based on the total citations, the top 3 are *Georgia Institute of Technology* from the USA (550 citations and 110 ACP), *Microsoft Research Asia* from China (115 citations and 16.43 ACP), and *National Cheng Kung University* from Taiwan (62 citations and 12.4 ACP). Ranking based on the ACP indicator yields the same result.

Table 5: The most productive authors in the NLP empowered mobile computing research field.

Rank	Name	Country	TP	TC	ACP	*H*	T100	*F*	FP	LP	CP
1	*Chen, Tao*	SG	4	108	27.00	4	3	4	3	0	4
1	*Mizzaro, Stefano*	IT	4	45	11.25	3	1	0	0	0	4
2	*Baek, Jin-Wook*	KR	3	27	9.00	3	1	0	1	0	3
2	*Bertino, Elisa*	USA	3	40	13.33	3	1	3	0	2	3
2	*Cacciapuoti, Angela Sara*	IT	3	18	6.00	2	1	4	3	0	3
2	*Caleffi, Marcello*	IT	3	18	6.00	2	1	4	0	2	3
2	*Christodoulakis, Stavros*	GR	3	9	3.00	1	0	0	0	3	3
2	*Crestani, F*	UK	3	37	12.33	2	1	0	0	3	3
2	*Jung, Jason J*	KR	3	4	1.33	1	0	2	0	2	3
2	*Karanastasi, Anastasia*	GR	3	9	3.00	1	0	0	1	0	3
2	*Kazasis, Fotis G*	GR	3	9	3.00	1	0	0	0	0	3
2	*Kim, Harksoo*	KR	3	4	1.33	1	0	7	0	2	3
2	*Lee, Chin-Hui*	USA	3	173	57.67	3	3	4	1	2	2
2	*Liu, Jia*	CN	3	3	1.00	1	0	6	0	1	3
2	*Muneyasu, Mitsuji*	JP	3	4	1.33	2	0	2	1	2	3
2	*Pierre, Samuel*	CA	3	38	12.67	2	1	0	0	0	3
2	*Seide, Frank*	CN	3	48	16.00	2	2	0	0	0	2
2	*Xie, Xing*	CN	3	51	17.00	3	2	3	1	0	3
2	*Yan, Yonghong*	CN	3	9	3.00	2	0	3	0	3	3
2	*Yeom, Heon Y*	KR	3	27	9.00	3	1	0	0	3	3

Abbreviations. CA: Canada; USA: the USA; UK: England; CN: China; KR: South Korea; GR: Greece; IT: Italy; JP: Japan; SG: Singapore; TP: total publications; TC: total citations; ACP: average number of citations per publication; *H*: *H*-index; T100: number of publications in the top 100 highly cited publications; *F*: number of publications with funding; FP: number of publications as first author; LP: number of publications as last author; CP: number of collaborated publications.

Table 6: The most productive affiliations in the NLP empowered mobile computing research field.

Rank	Name	Country	TP	TC	ACP	*H*	T100	*F*	FP	CP
1	*Nanyang Technological University*	SG	8	87	10.88	5	3	1	4	5
1	*Tsinghua University*	CN	8	42	5.25	4	1	21	8	4
3	*Microsoft Research Asia*	CN	7	115	16.43	5	4	3	3	6
3	*National Taiwan University*	TW	7	83	11.86	5	3	3	5	4
5	*Georgia Institute of Technology*	USA	5	550	110.00	4	4	6	2	4
5	*Massachusetts Institute of Technology*	USA	5	10	2.00	2	0	9	4	4
5	*National Cheng Kung University*	TW	5	62	12.40	3	2	6	4	1
5	*Purdue University*	USA	5	47	9.40	4	1	4	1	5
9	*Indian Institute of Technology*	IN	4	35	8.75	3	1	3	4	1
9	*Microsoft Corporation*	USA	4	28	7.00	3	1	1	4	2
9	*The Pennsylvania State University*	USA	4	26	6.50	3	0	8	2	2
9	*Seoul National University*	KR	4	31	7.75	4	1	5	4	0
9	*University of Strathclyde*	UK	4	43	10.75	2	1	0	4	0
9	*University of Udine*	IT	4	45	11.25	3	1	0	1	3
9	*Zhejiang University*	CN	4	43	10.75	2	1	5	3	3

Abbreviations. USA: the USA; UK: England; CN: China; SG: Singapore; TW: Taiwan; IN: India; KR: South Korea; IT: Italy; TP: total publications; TC: total citations; ACP: average number of citations per publication; *H*: *H*-index; T100: number of publications in the top 100 highly cited publications; *F*: number of publications with funding; FP: number of publications as first affiliation; CP: number of collaborated publications.

3.5. Geographical Distribution. The 471 publications are from 60 countries/regions. The number of publications affiliated with 1 country/region range [61, 105), 3 countries/regions range [37, 61), and 5 range [11, 17). Table 7 shows top 15 most productive countries/regions in the field. Figure 4 illustrates geographical distributions of the publications. The top 4 countries are the USA (105 publications, 1,795 citations, 17.1 ACP, and 22 *H*-index), China (61 publications, 372 citations, 6.1 ACP, and 10 *H*-index), England (44 publications, 418 citations, 9.5 ACP, and 12 *H*-index), and South Korea (41 publications, 281 citations, 6.85 ACP, and 8 *H*-index). Among the 105 publications from the USA, 32 appear in the list of top

Table 7: The most productive countries/regions in the NLP empowered mobile computing research field.

Rank	Country	TP	TC	ACP	*H*	T100	FP (%)	Single-country/region		International collaboration	
								ACP	TP (%)	ACP	TFC (*n*)
1	USA	105	1,795	17.10	22	32	77.14	20.78	60.00	11.57	CN (12)
2	CN	61	372	6.10	10	10	91.80	4.17	57.38	9.04	USA (12)
3	UK	44	418	9.50	12	11	61.36	11.68	63.64	5.69	IE/CH (2)
4	KR	41	281	6.85	8	6	92.68	7.03	85.37	5.83	CN/USA (3)
5	TW	37	399	10.78	11	11	94.59	11.07	81.08	9.57	USA (4)
6	JP	24	77	3.21	3	1	79.17	1.44	75.00	8.50	CN (3)
7	IT	21	299	14.24	10	9	80.95	13.19	76.19	17.60	USA (3)
8	AU	18	218	12.11	7	7	61.11	18.00	38.89	8.36	USA (5)
8	CA	18	313	17.39	9	4	88.89	20.38	72.22	9.60	N/A
10	FR	17	157	9.24	6	5	64.71	4.45	64.71	18.00	CN/USA (2)
10	GR	17	38	2.24	3	0	100.00	2.24	100.00	0.00	N/A
10	ES	17	124	7.29	7	2	88.24	6.43	82.35	11.33	USA (2)
13	SG	16	355	22.19	9	7	75.00	14.90	62.50	34.33	USA (2)
14	HK SAR	15	98	6.53	6	2	53.33	9.17	40.00	4.78	CN/USA (4)
15	DE	14	114	8.14	5	3	85.71	8.11	64.29	8.20	CN (12)

Abbreviations. USA: America; UK: England; CN: China; SG: Singapore; TW: Taiwan; KR: South Korea; IT: Italy; JP: Japan; AU: Australia; CA: Canada; FR: France; GR: Greece; ES: Spain; HK SAR: Hong Kong SAR; DE: Germany; IE: Ireland; CH: Switzerland; TP: total publications; TC: total citations; ACP: average number of citations per publication; *H*: *H*-index; T100: number of publications in the top 100 highly cited publications; FP (%): percentage of publications as first affiliation; TFC (*n*): number of cooperation times with the closest collaborator, where $n \geq 2$.

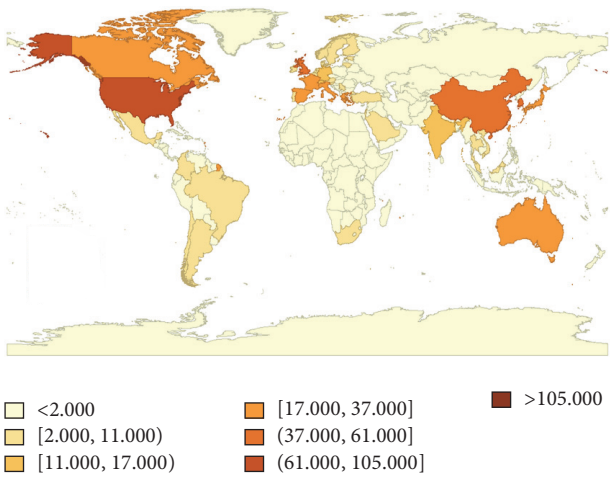


Figure 4: Geographical distributions of the NLP empowered mobile computing research publications.

100 most influential publications. It is noted that publications from Singapore have the highest ACP, which indicates the high quality of the publications. As for most of the top 15 productive countries/regions, the international collaboration rates are around 30%, except for Greece with 0 and Australia with 61.11%. The USA is the closest collaborator for 9 of the 15 countries/regions. The ACP of internationally collaborated publications is much higher than that of noninternationally collaborated publications for countries/regions like China, Japan, Italy, France, Spain, and Singapore. This potentially indicates that international collaboration can improve the quality of their publications.

Since the publications are mainly distributed in the USA, China, England, and South Korea, we further explored the annual publication distributions for these 4 countries, as shown in Figure 5. The number of publications for the USA and China is on the whole presenting upward trend in fluctuation. As for the USA, the number increases from 2 in 2000 to 9 in 2007 but dwindles to 2 in 2010. After that, the upward trend becomes more significant. The situation for China is quite like that for the USA after 2010, witnessing the great mass upsurge on the NLP empowered mobile computing research in these two countries since 2010. As for England and South Korea, the number of publications does not increase much in fluctuation with years going on.

3.6. Cooperation Relationship. Figure 6 shows the trends of the international collaborative and the percentage of international collaborative publications. We found that the international collaborative publications increase during the years 2000–2016. The percentage of international collaborations increases from 8.33% in 2000 to 32.73% in 2016. This indicates that international collaborations in the NLP empowered mobile computing research field have become increasingly important.

Figures 7 and 8 present the institutional level of cooperation and the author level of cooperation, respectively. The cooperation between different institutions is becoming more and more frequent. The percentage of institution-collaborative publication increases from 16.67% in 2000 to 58.18% in 2016. More than 90% of the publications are multiauthored since 2011. It is worth noticing that the percentage reaches up to 100% in 2015.

Furthermore, the cooperation relations for specific countries/regions, affiliations, and authors were visualized with

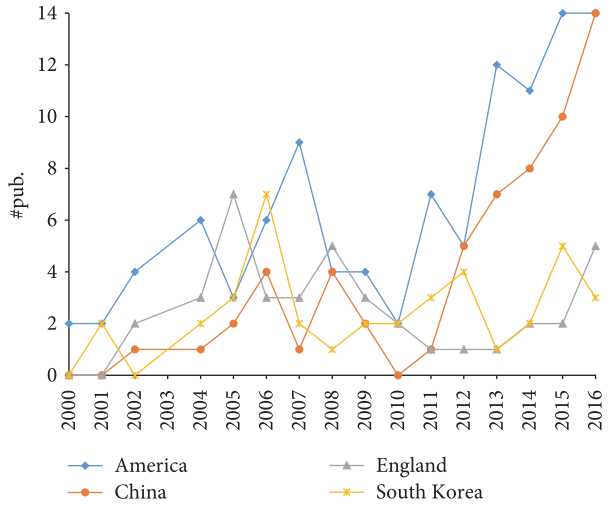


Figure 5: Publication distributions by year for the top 4 countries/regions.

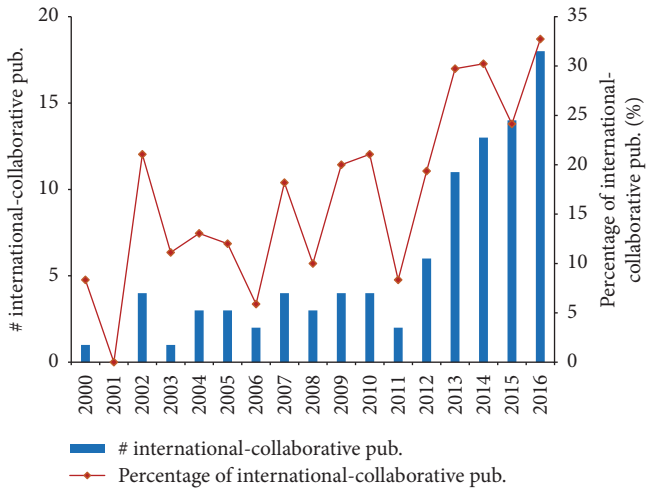


Figure 6: International collaborative publication distribution by year.

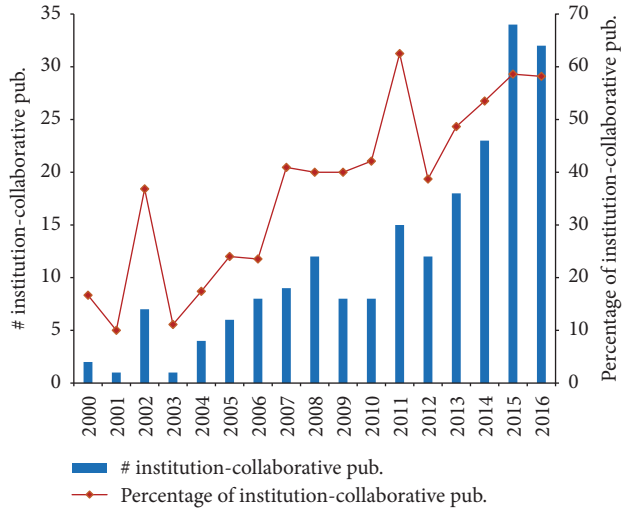


Figure 7: Institution-collaborative publication distribution by year.

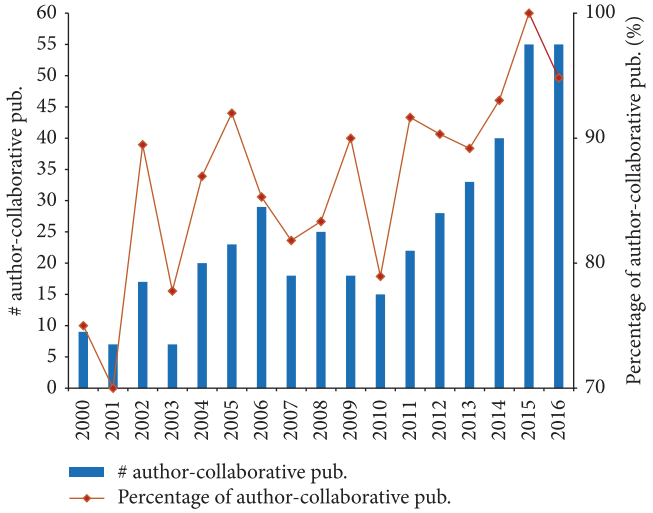


Figure 8: Author-collaborative publication distribution by year.

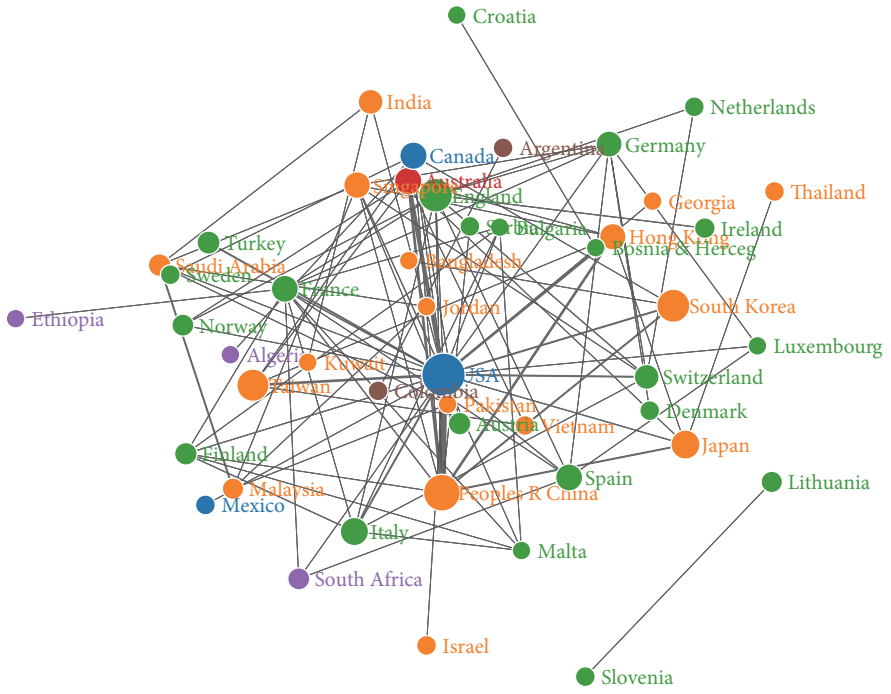


Figure 9: Cooperation network of 48 countries/regions (node colors represent different continents, e.g., orange for Asia, blue for North America, green for Europe, red for Oceania, purple for Africa, and brown for South America). The network can be accessed via the link (http://www.zhukun.org/haoty/resources.asp?id=NLPEMC_cocountry).

social network analysis. A cooperation network for 48 countries/regions is shown in Figure 9. 17 of them come from Asia (represented as orange nodes), 3 from North America (represented as blue nodes), 22 from Europe (represented as green nodes), 3 from Africa (represented as purple nodes), 2 from South America (represented as brown nodes), and 1 from Oceania (represented as red node). There are 141 affiliations with the number of publications ≥ 2, and there exists cooperation among 91 of them. Figure 10 shows a cooperation network of the 91 affiliations. 23 of the 91 affiliations are from the USA and 14 from China. As for cooperation of author level, there are 98 authors with publication count ≥ 2. among them, 65 authors involve in cooperation. We created a cooperation network of the 65 authors, as shown in Figure 11.

3.7. Topic Discovery and Distribution. By setting TF-IDF value threshold as 0.1, the terms were ranked by frequency. Table 8 lists top 20 most frequent terms, in which the top 5

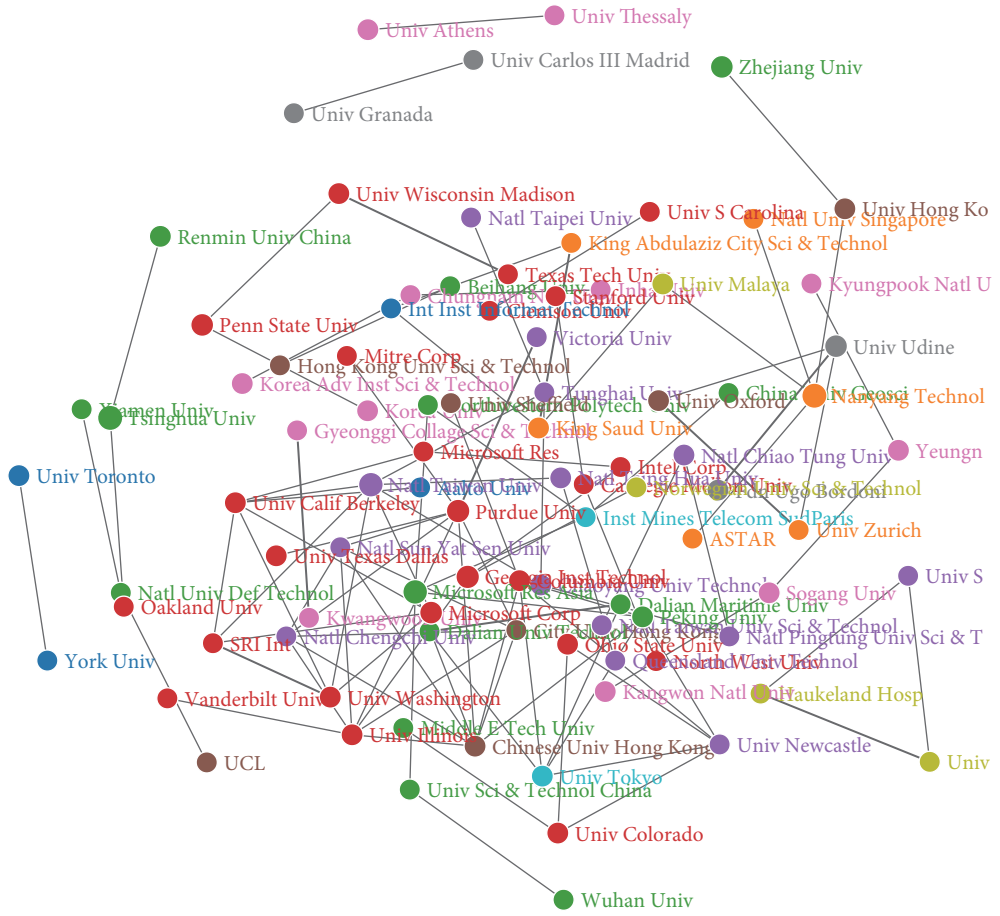


FIGURE 10: Cooperation network of 91 affiliations (node colors represent different countries/regions, e.g., red for the USA, pink for South Korea, and purple for Australia). The network can be accessed via the link (http://www.zhukun.org/haoty/resources.asp?id=NLPEMC_coaffiliation).

terms are "Agent" (369), "Image" (215), "Sentiment" (128), "Dialogue" (83), and "Health" (81). Figure 12 presents the perplexities of models fitted by using Gibbs sampling with different numbers of topics. The result suggests that the optimal topic number is between 40 and 80. Hence, we set the topic number as 40. The α was set to the mean value 0.01101332 in the cross-validation fitted using VEM. Using the parameters, we estimated the LDA model using Gibbs sampling. By semantics analysis of representative terms in each topic, as well as reviewing text intention of the corresponding publications, we assigned potential theme to each topic. The order of topics are determined based on Hellinger distance. Specifically, Topic 36 is the best matching topic and Topic 11 ranks 2nd, while Topic 37 is the worse matching one. Due to space limitation, Table 9 only displays the top 10 best matching topics with the most frequent terms. Each publication was assigned to the most likely topic with the highest posterior probability. Integrating topic proportions for all the publications, we obtained a topic distribution. The 4 most frequent research topics are Topic 36 (6.38%), Topic 4 (4.26%), Topic 11 (3.83%), and Topic 17 (3.83%), while the 4 least frequent research topics are Topic 26 (1.49%), Topic 23 (1.28%), Topic 10 (1.06%), and Topic 20 (1.06%).

We used the AP clustering analysis to perform the cluster analysis of the 40 topics. One way for measuring topic similarity is based on term-level similarity with the hypothesis that topics may contain the same terms. The clustering result based on term-topic posterior probability matrix is shown in Figure 13, where the 40 topics are categorized into 8 groups.

Identifying emerging research topics can provide valuable insights into the development of the research field. Likewise, identification of fading research topics can also help understand the hot spots evolution [40]. We then explored the annual publication proportions of the 40 research topics, as shown in Figure 14. We used Mann–Kendall test [41], a nonparametric trend test, to examine whether increasing or decreasing trends are existing in the 40 topics. Test results show that 12 topics, including Topic 1, Topic 4, Topic 7, Topic 10, Topic 14, Topic 18, Topic 20, Topic 26, Topic 29, Topic 32, Topic 33, and Topic 39, present a statistically significant increasing trend. While Topic 36 presents a statistically significant decreasing trend, both at the two-sided $p = 0.05$ levels.

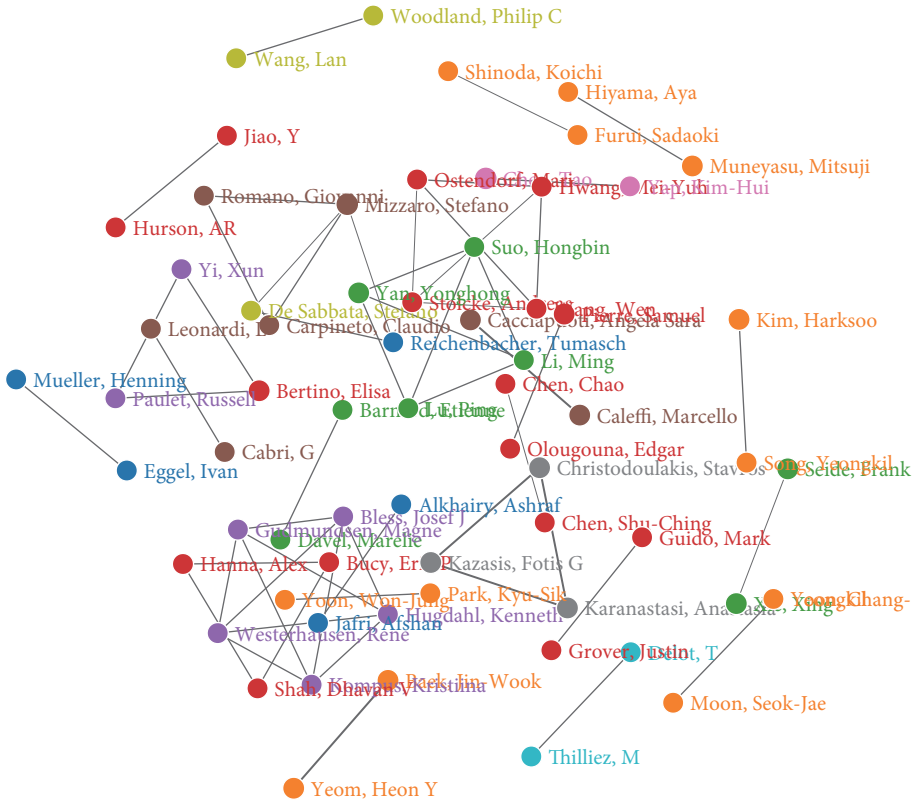


FIGURE 11: Cooperation network of 65 authors (node colors represent different countries/regions, e.g., range for South Korea, red for the USA, purple for Australia, green for China, and brown for Italy). The network can be accessed via the link (http://www.zhukun.org/haoty/resources.asp?id=NLPEMC_coauthor).

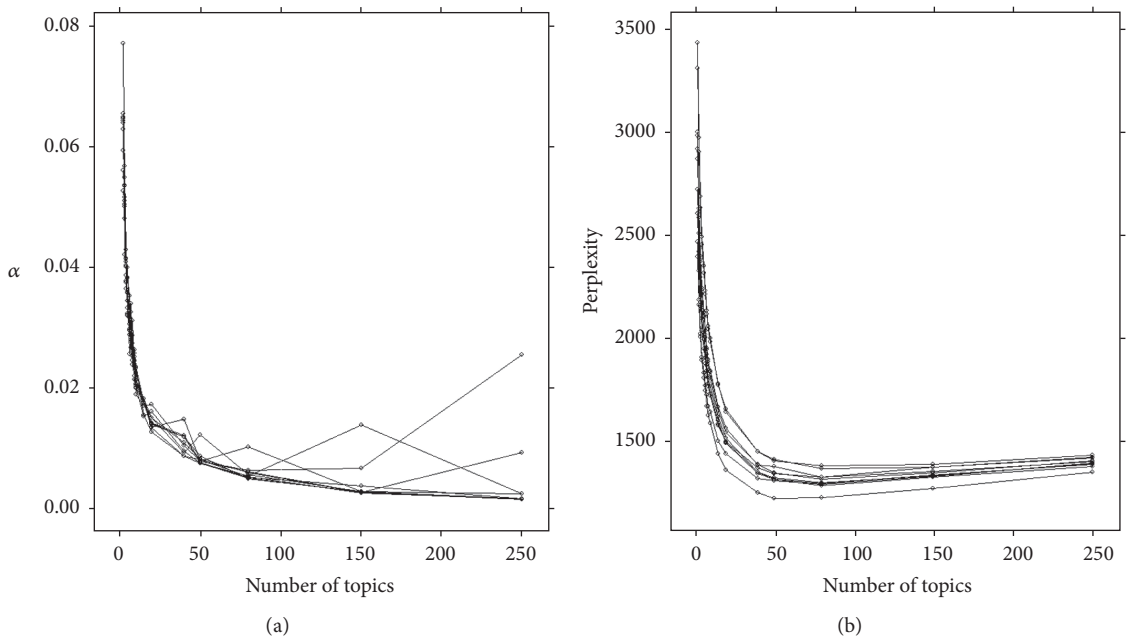


FIGURE 12: (a) Estimated α value for the models fitted using VEM. (b) Perplexities of the test data for the models fitted by using Gibbs sampling. Each line corresponded to one of the folds in the 10-fold cross-validation.

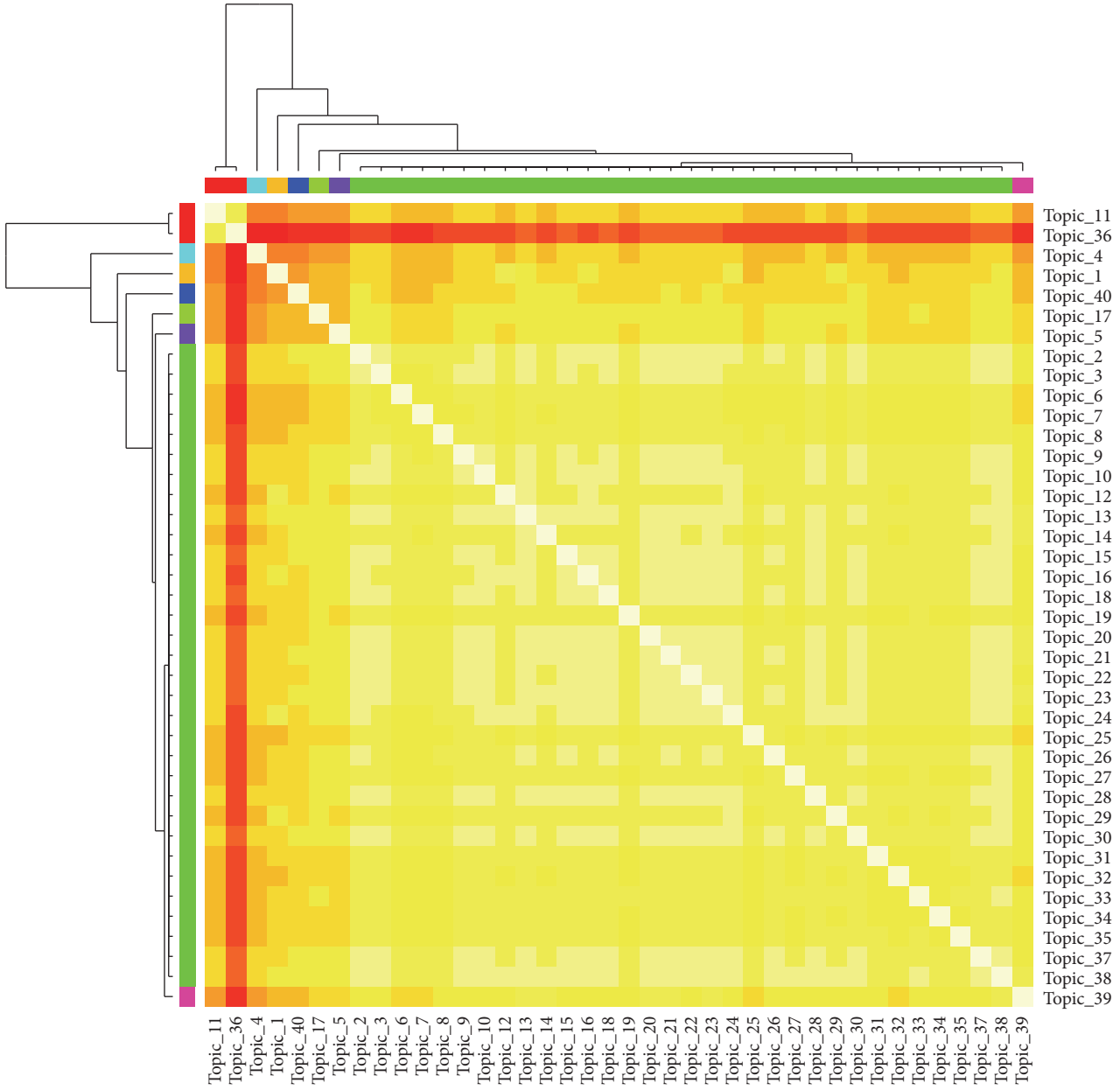


FIGURE 13: The visualized result of hierarchical clustering based on term-topic posterior probability matrix.

4. Discussions

This study provides a most up-to-date bibliometric analysis on the publications in WoS during the years 2000–2016 in the NLP empowered mobile computing research field. Some interesting findings are discussed below.

The annual number of the publication distribution shows a significant growth trend, from 12 publications in 2000 to 55 publications in 2016. This indicates a growing interest in the research field.

The literature characteristics analysis shows that the 417 publications are widely dispersed throughout 287 journals. 11 most productive journals together contribute about 21% of the total publications. The top 3 are *IEEE/ACM Transactions on Audio Speech and Language Processing*, *Speech Communication*, and *Computer Speech and Language*. *Computer science* is the most shared subject among these 11 journals. Journal *Information Sciences* possesses the highest IF, SJR, 5-Year IF, and CiteScore, except for the SNIP score in year 2016.

Top 3 most influential publications are: [35] by Miao et al. published in 2010, [36] by MacKenzie and Soukoreff published in 2002, and [37] by Strayer and Drews published in 2007.

There are 1,408 authors and 544 affiliations involved in the publications. Most authors (79.18%) have only 1 publication, and 4.25% of the authors have 3 or more publications. The most productive authors are *Chen, Tao* from Singapore and *Mizzaro, Stefano* from Italy. In addition, most affiliations

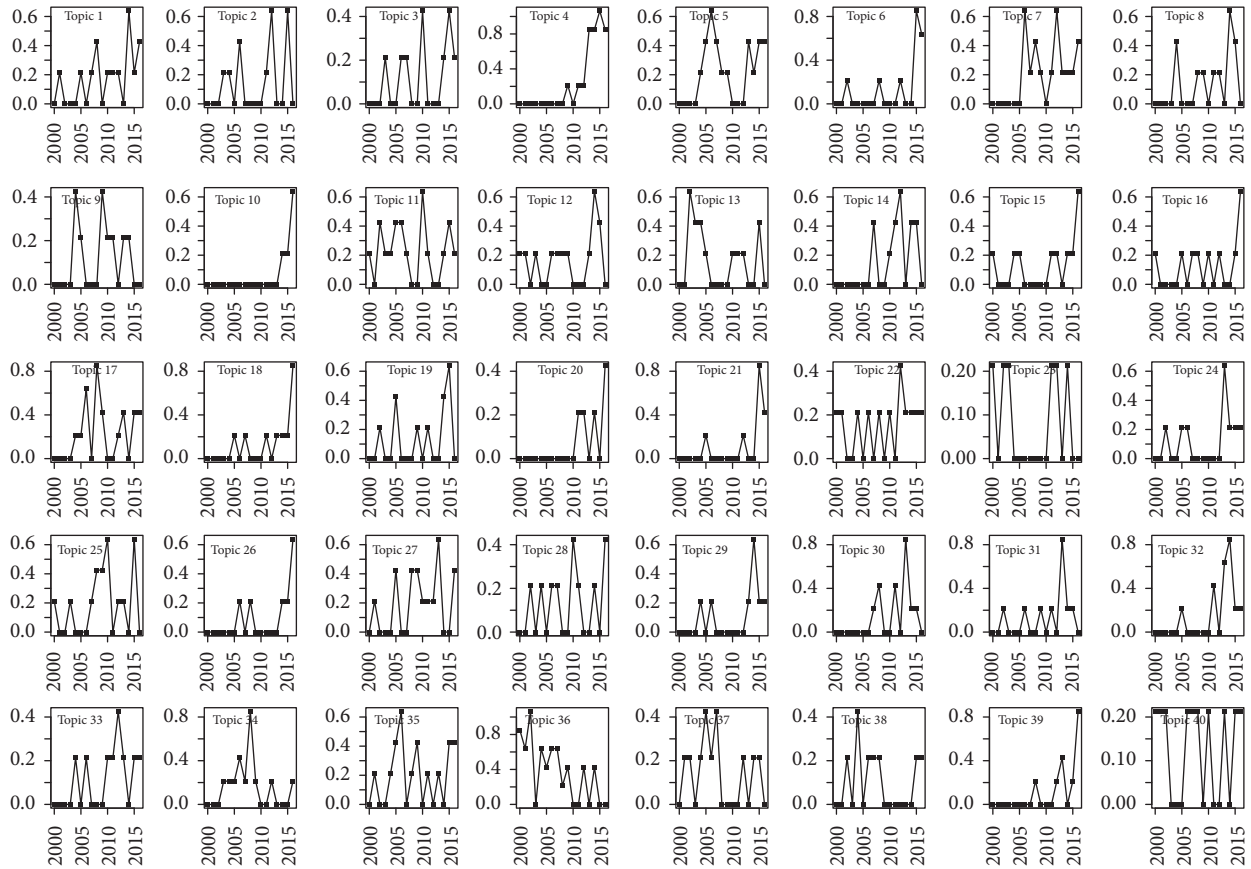


Figure 14: The trends of the 40 research topics during 2000–2016 (x-coordinate as year, y-coordinate as proportion %).

Table 8: Top 20 most frequent terms.

Rank	Stemmed terms	Occurrence number		
		Total	2000–2008	2009–2016
1	Agent	369	250	119
2	Image	215	70	145
3	Sentiment	128	0	128
4	Dialogue	83	49	34
5	Health	81	2	79
6	Music	76	27	49
7	Radio	74	10	64
8	Unit	74	51	23
9	Adaptation	70	40	30
10	Relevance	69	29	40
11	Geographic	66	37	29
12	Short Messages	66	9	57
13	Protocol	65	20	45
14	Chinese	64	29	35
15	Medical	60	16	44
16	Recommendation	60	4	56
17	Clustering	54	20	34
18	Privacy	54	9	45
19	Ad hoc	53	9	44
20	Traffic	52	17	35

(70.06%) have 1 publication. 11.89% of the affiliations have 3 or more publications. The most productive affiliations are *Nanyang Technological University* from Singapore and *Tsinghua University* from China. *Lee, Chin-Hui* from USA with 57.67 ACP ranks 1st among top 20 productive authors, and *Georgia Institute of Technology* from USA with 110 ACP ranks 1st among 15 most productive affiliations.

Through geographic visualization analysis, 60 countries/regions have participated in the publications. The top 15 productive countries/regions are developed countries/regions, except for China. As the top 2, the USA and China have shown a significant growth in the numbers of scientific publications since 2010. These numbers are predicted to continue to increase in the coming years. This partially reflects the need of the development of NLP techniques in solving mobile computing issues.

Scientific collaboration analysis shows that there are significant growth of international collaborations, institution-collaborations as well as author-collaborations. Through social network analysis, we found that researchers tend to collaborate with others within the same country or area, with institutions under similar administration, or with a neighboring country or area. However, some research institutions might have separate administration arrangements from their associated universities or hospitals and a researcher might be affiliated with multiple institutions. The co-authors

Table 9: Top 15 most frequent terms for the top 10 best matching topics.

Topic	Potential theme	Top high frequency terms
36	Mobile agent computing	Agent; Coordination; Java; Migration; Protocol; Mobile-agent; Failure; Itinerary; Filtering; Turkish; Attack; Commerce; Context-aware; Truncation; Crash
11	Mobile agent computing	Agent; Planning; Ontology; Cloud; Multi-agent; Net; Interoperability; Neural; Peer-to-Peer; Broadband; Instruction; Complementarity; Natural Language; Traffic; Grounding
32	Mobile privacy and security	Privacy; Private; Secure; Location-Based Services; Encryption; Points of Interest; Protection; Approximate; Attack; Path; Privacy-preserving; Streaming; Password; Protocol; Cryptosystem
1	Image and syllable events	Image; Particular Allophones; Re-ranking; Composite Phoneme; Simple Phonemes; Syllable; Thing; iPad; On-Premise Signs; Spreading; Bow; Modern Orthography; Arabic; Content-based; Descriptor
4	Mobile social media computing	Sentiment; Opinion; Twitter; Tweet; Customer; Suggestion; Emojis; Emotion; Micro-blog; Protest; Brand; Suggestive; Microblog; Orientation; Box
8	Mobile radio	Radio; Phone-in; Localization; Australian; Formulation; Island; Reporting; Talkback; Involvement; Caller; Dialogic; Stance; Backlinking; Cloud; French
5	Mobile location computing	Geographic; Relevance; Seeking; Innovation; Subspace; Tourism; Birthright; Firm; Flier; Sensing; TILES (Temporal, Identity, Location, Environmental and Social); Cross-space; Location-aware; Personalized; Reposting
40	Context-aware computing	Dialogue; Context-aware; Estonian; Clarification; Array; Problematic; Reformulation; Verbose; Email; Mobile Information Services enabled by Mobile Publishing; Non-understanding; Publishing; Agent; Directive; Reinforcement
10	Second screen response	Gesture; Debate; PreFrontal Cortex; Adult; Presidential; Walking; Facial; Twitter; Educational; Gait; Political; Touch; Biometrics; Blink; Cortex
35	Language learning and modeling	Chinese; Information Retrieval; Peer-to-Peer; Conditional Random Field; Update; Apprentice; Affordances; Disyllabic; Website; Workplace; Self-study; Skip-chain; Descriptive; Mobile Peer-to-Peer; Multilingual

might actually work together but are affiliated with different institutions. Therefore, it is worth noticing that institution-wise collaboration might not be the actual collaboration among institutions.

Most topics identified using LDA method are recognizable, as they are related to major issues in the research field. Due to space constraints, here we only provide interpretations of some representative topics.

Topic 36 and Topic 11 contain words such as "Agent", "Mobile-agent", "Multi-agent", "Itinerary", "Migration", "Protocol", and "Truncation". Thus, Topic 36 and Topic 11 pertain to *mobile agent computing*. As an emerging and exciting paradigm for mobile computing applications [42], mobile agent can not only support mobile computers and disconnected operations but also provide an efficient, convenient and robust programming paradigm for implementing distributed applications. The use of mobile agent can bring about significant benefits, e.g., reduction of network traffic, overcoming network latency, and seamless system integration. Therefore, mobile agent is well adapted to the domain of mobile computing.

Topic 32 discusses *events about mobile privacy and security*. Words in this topic include "Privacy", "Private", "Secure", "Encryption", "Privacy-preserving", "Password", and "Cryptosystem". As pointed out by Mollah et al. [43], security and privacy challenges are introduced with the development of mobile cloud computing which aims at relieving challenges of the resource constrained mobile devices in mobile computing area. Studies centering on mobile privacy can be found. For example, Xi et al. [44] applied Private Information Retrieval techniques in finding the shortest path between an origin and a destination in location privacy issues without the risk of disclosing their privacy.

Topic 1 discusses *mobile computing on image and syllable events*. It includes words such as "Image", "Syllable", "Re-ranking", "Content-based", "Composite Phoneme", "Simple Phonemes", and "Modern Orthography". Image search in mobile device is quite worthy of challenge [45]. Many researchers are seeking ways to solve this problem. For example, Cai et al. [46] presented a new geometric reranking algorithm specific for small vocabulary in aforementioned scenarios based on Bag-of-Words model for image retrieval. Mobile computing on syllable events is another focus. A representative work is by Eddington and Elzinga [47]. They conducted a quantitative analysis on the phonetic context of word-internal flapping with great attention paid to stress placement, following phone, and syllabification.

Topic 4 mainly focuses on *mobile social media event*. Words like "Twitter", "Sentiment", "Tweet", "Emojis", "Micro-blog", "Opinion", "Public", and "Emotion" can be found within this topic. With the rapid development of social network, information spreading and evolution is facilitated with popularity of the environment of wireless communication, especially social media platform on mobile terminals [48]. Researchers are gradually paying attention to this area. For example, based on 100 million collected

messages from Twitter, Wang et al. [49] presented a hybrid model for sentimental entity.

Based on topic distributions, we found that *mobile agent computing*, *mobile social media computing*, and *sound related event computing* are 3 highest-frequent research themes. From Figure 14 as well as Mann–Kendall test results, we found that some research themes present a statistically significant increasing trend, e.g., *image and syllable related events*, *mobile social media computing*, and *healthy related events*, while researches on *mobile agent computing* presents a statistically significant decreasing trend.

In the thematic analysis, the optimal number of topics was selected as 40 by a statistical measure of model fitting the data. However, mechanical reliance on statistical measures might lead to the selection of a less meaningful topic model [50]. Hence, we manually checked the robustness of the results by confirming identified topics using a qualitative assessment with the basis of prior knowledge. For each topic, we checked the semantic coherence of its high-frequency terms and examined the contents of publication with a high proportion of this topic.

Through the AP clustering analysis on the 40-topics, 8 clusters were identified, i.e., *mobile agent computing, mobile social media computing, image and syllable related events, context-aware computing, sound related events, mobile location computing, healthy related events*, and other events. The results of AP clustering analysis are on the whole sensible and easy-to-understand. However, we still found that the 8 categories vary a lot in topic numbers. One possible reason is the choice of clustering method. We then adopted hierarchical clustering method with category number setting to 8. The result was similar with AP clustering. Another possible reason is the sample size since the number of the relevant publications in WoS is limited.

This study is the first to thoroughly explore research status of the NLP empowered mobile computing research field in the statistical perspective. The study provides a comprehensive overview and an intellectual structure of the field from 2000 to 2016. The findings can potentially help researchers especially newcomers systematically understand the development of the field, learn the most influential journals, recognize potentially academic collaborators, and trace research hotspots.

For future work, there are several directions. First, more comprehensive data is expected to be included. Though WoS is a widely applied repository for bibliometric analysis due to its high authority, some relevant conference proceedings have not been indexed yet in WoS. Second, we intend to employ different data clustering methods and compare clustering results for deeper cluster analyzing.

5. Conclusions

We conducted a bibliometric analysis on natural language processing empowered mobile computing research publications from Web of Science published during years 2000–2016. The literature characteristics were uncovered using a descriptive statistics method. Geographical publication distribution was explored using a geographic visualization method. By applying a social network analysis method, cooperation relationships among countries/regions, affiliations, and authors were displayed. Finally, topic discovery and distribution were presented using a LDA method and an AP clustering method. We believe the analysis can help researchers comprehend the collaboration patterns and distribution of scholarly resources and research hot spots in the research field more systematically.

Disclosure

Tianyong Hao and Yi Zhou are the corresponding authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The work was substantially supported by the grant from National Natural Science Foundation of China (no. 61772146), the Innovative School Project in Higher Education of Guangdong Province (No. YQ2015062), Science and Technology Program of Guangzhou (no. 201604016136), and Major Project of Frontier and Key Technical Innovation of Guangdong Province (no. 2014B010118003).

References

[1] G. Deepak and B. S. Pradeep, "Challenging issues and limitations of mobile computing," vol. 3, pp. 177–181, 2012.

[2] K.-Y. Chung, J. Yoo, and K. J. Kim, "Recent trends on mobile computing and future networks," *Personal and Ubiquitous Computing*, vol. 18, no. 3, pp. 489–491, 2014.

[3] M. Chen, J. Pan, Q. Zhao, and Y. Yan, "Multi-task learning in deep neural networks for Mandarin-english code-mixing speech recognition," *IEICE Transaction on Information and Systems*, vol. E99D, no. 10, pp. 2554–2557, 2016.

[4] N. Ilayaraja, F. Mary Magdalene Jane, M. Safar, and R. Nadarajan, "WARM Based Data Pre-fetching and Cache Replacement Strategies for Location Dependent Information System in Wireless Environment," *Wireless Personal Communications*, vol. 90, no. 4, pp. 1811–1842, 2016.

[5] L.-H. Wong, R. B. King, C. S. Chai, and M. Liu, "Seamlessly learning Chinese: contextual meaning making and vocabulary growth in a seamless Chinese as a second language learning environment," *Instructional Science*, vol. 44, no. 5, pp. 399–422, 2016.

[6] O. J. Räsänen and J. P. Saarinen, "Sequence prediction with sparse distributed hyperdimensional coding applied to the analysis of mobile phone use patterns," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 9, pp. 1878–1889, 2016.

[7] M. Puppala, T. He, S. Chen et al., "METEOR: An Enterprise Health Informatics Environment to Support Evidence-Based Medicine," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 12, pp. 2776–2786, 2015.

[8] A. O. Adesina, K. K. Agbele, A. P. Abidoye, and H. O. Nyongesa, "Text messaging and retrieval techniques for a mobile health

information system," *Journal of Information Science*, vol. 40, no. 6, pp. 736–748, 2014.

[9] W.-T. Chiu and Y.-S. Ho, "Bibliometric analysis of tsunami research," *Scientometrics*, vol. 73, no. 1, pp. 3–17, 2007.

[10] J. M. Merigó, A. M. Gil-Lafuente, and R. R. Yager, "An overview of fuzzy research with bibliometric indicators," *Applied Soft Computing*, vol. 27, pp. 420–433, 2015.

[11] D. Bouyssou and T. Marchant, "Ranking scientists and departments in a consistent manner," *Journal of the Association for Information Science and Technology*, vol. 62, no. 9, pp. 1761–1769, 2011.

[12] A. Mazloumian, "Predicting Scholars' Scientific Impact," *PLoS ONE*, vol. 7, no. 11, Article ID e49246, 2012.

[13] X. Chen, H. Xie, F. Wang, Z. Liu, J. Xu, and T. Hao, "Natural Language Processing in Medical Research: A Bibliometric Analysis," *BMC Medical Informatics and Decision Making*, vol. 18, supplement 1, no. 14, 2018.

[14] Y. Geng, W. Chen, Z. Liu et al., "A bibliometric review: Energy consumption and greenhouse gas emissions in the residential sector," *Journal of Cleaner Production*, vol. 159, pp. 301–316, 2017.

[15] A. Khan, N. Choudhury, S. Uddin, L. Hossain, and L. A. Baur, "Longitudinal trends in global obesity research and collaboration: A review using bibliometric metadata," *Obesity Reviews*, vol. 17, no. 4, pp. 377–385, 2016.

[16] N. Roig-Tierno, T. F. Gonzalez-Cruz, and J. Llopis-Martinez, "An overview of qualitative comparative analysis: A bibliometric analysis," *Journal of Innovation Knowledge*, vol. 2, no. 1, pp. 15–23, 2017.

[17] G. Albort-Morant and D. Ribeiro-Soriano, "A bibliometric analysis of international impact of business incubators," *Journal of Business Research*, vol. 69, no. 5, pp. 1775–1779, 2016.

[18] J. M. Merigó and J.-B. Yang, "A bibliometric analysis of operations research and management science," *OMEGA - The International Journal of Management Science*, vol. 73, pp. 37–48, 2017.

[19] K. Zhang, Q. Wang, Q.-M. Liang, and H. Chen, "A bibliometric analysis of research on carbon tax from 1989 to 2014," *Renewable & Sustainable Energy Reviews*, vol. 58, pp. 297–310, 2016.

[20] K. Randhawa, R. Wilden, and J. Hohberger, "A Bibliometric Review of Open Innovation: Setting a Research Agenda," *Journal of Product Innovation Management*, vol. 33, no. 6, pp. 750–772, 2016.

[21] A. Yataganbaba and I. Kurtbaş, "A scientific approach with bibliometric analysis related to brick and tile drying: A review," *Renewable & Sustainable Energy Reviews*, vol. 59, pp. 206–224, 2016.

[22] X. Chen, B. Chen, C. Zhang, and T. Hao, "Discovering the Recent Research in Natural Language Processing Field Based on a Statistical Approach," in *Emerging Technologies for Education*, vol. 10676 of *Lecture Notes in Computer Science*, pp. 507–517, Springer International Publishing, Cham, 2017.

[23] H. J. Kim, D. Y. Yoon, E. S. Kim, K. Lee, J. S. Bae, and J.-H. Lee, "The 100 most-cited articles in neuroimaging: A bibliometric analysis," *Results in Physics*, vol. 139, pp. 149–156, 2016.

[24] X. Chen, H. Weng, and T. Hao, "A Data-Driven Approach for Discovering the Recent Research Status of Diabetes in China," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 10594, pp. 89–101, 2017.

[25] W. R. Tobler, "A computer movie simulating urban growth in the detroit region," *Economic Geography*, vol. 46, supplement 1, pp. 234–240, 1970.

[26] E. Otte and R. Rousseau, "Social network analysis: a powerful strategy, also for the information sciences," *Journal of Information Science*, vol. 28, no. 6, pp. 441–453, 2002.

[27] D. R. Radev, M. T. Joseph, B. Gibson, and P. Muthukrishnan, "A bibliometric and network analysis of the field of computational linguistics," *Journal of the Association for Information Science and Technology*, vol. 67, no. 3, pp. 683–706, 2016.

[28] M. Grandjean, "A social network analysis of Twitter: Mapping the digital humanities community," *Cogent Arts and Humanities*, vol. 3, no. 1, Article ID 1171458, 2016.

[29] M. Salathé, M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones, "A high-resolution human contact network for infectious disease transmission," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 107, no. 51, pp. 22020–22025, 2010.

[30] J. Scott, "Social network analysis," *Sage*, 2017.

[31] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 4-5, pp. 993–1022, 2003.

[32] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by Gibbs sampling," in *Proceedings of the the 43rd Annual Meeting*, pp. 363–370, Ann Arbor, Michigan, June 2005.

[33] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *American Association for the Advancement of Science: Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[34] A. F. El-Samak and W. Ashour, "Optimization of Traveling Salesman Problem Using Affinity Propagation Clustering and Genetic Algorithm," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 5, no. 4, 2015.

[35] G. Miao, N. Himayat, and G. Y. Li, "Energy-efficient link adaptation in frequency-selective channels," *IEEE Transactions on Communications*, vol. 58, no. 2, pp. 545–554, 2010.

[36] I. S. MacKenzie and R. W. Soukoreff, "Text entry for mobile computing: Models and methods, theory and practice," *Human–Computer Interaction*, vol. 17, no. 2-3, pp. 147–198, 2002.

[37] D. L. Strayer and F. A. Drews, "Cell-phone-induced driver distraction," *Current Directions in Psychological Science*, vol. 16, no. 3, pp. 128–131, 2007.

[38] J. Cao, T. Chen, and J. Fan, "Landmark recognition with compact BoW histogram and ensemble ELM," *Multimedia Tools and Applications*, 2015.

[39] M. M. Mostafa, "More than words: social networks' text mining for consumer brand sentiments," *Expert Systems with Applications*, vol. 40, no. 10, pp. 4241–4251, 2013.

[40] H. Jiang, M. Qiang, and P. Lin, "A topic modeling based bibliometric exploration of hydropower research," *Renewable & Sustainable Energy Reviews*, vol. 57, pp. 226–237, 2016.

[41] H. B. Mann, "Nonparametric tests against trend," *Econometrica*, vol. 13, pp. 245–259, 1945.

[42] D. B. Lange and M. Oshima, "Seven Good Reasons for Mobile Agents," *Communications of the ACM*, vol. 42, no. 3, pp. 88-89, 1999.

[43] M. B. Mollah, M. A. K. Azad, and A. Vasilakos, "Security and privacy challenges in mobile cloud computing: Survey and way ahead," *Journal of Network and Computer Applications*, vol. 84, pp. 38–54, 2017.

[44] Y. Xi, L. Schwiebert, and W. Shi, "Privacy preserving shortest path routing with an application to navigation," *Pervasive and Mobile Computing*, vol. 13, pp. 142–149, 2014.

[45] T. Yan, V. Kumar, and D. Ganesan, "CrowdSearch: exploiting crowds for accurate real-time image search on mobile phones," in *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services (MobiSys '10)*, pp. 77–90, ACM, San Francisco, Calif, USA, June 2010.

[46] Y. Cai, S. Li, Y. Cheng, and R. Ji, "Local consistent hierarchical Hough Match for image re-ranking," *Journal of Visual Communication and Image Representation*, vol. 37, pp. 32–39, 2016.

[47] D. Eddington and D. Elzinga, "The phonetic context of american english flapping: Quantitative evidence," *Language and Speech*, vol. 51, no. 3, pp. 245–266, 2008.

[48] X. Wang, H. Zhang, S. Yuan, J. Wang, and Y. Zhou, "Sentiment processing of social media information from both wireless and wired network," *EURASIP Journal on Wireless Communications and Networking*, vol. 2016, no. 1, article no. 164, 2016.

[49] Z. Wang, X. Cui, L. Gao, Q. Yin, L. Ke, and S. Zhang, "A hybrid model of sentimental entity recognition on mobile social media," *EURASIP Journal on Wireless Communications and Networking*, vol. 2016, no. 1, article no. 253, 2016.

[50] K. E. C. Levy and M. Franklin, "Driving Regulation: Using Topic Models to Examine Political Contention in the U.S. Trucking Industry," *Social Science Computer Review*, vol. 32, no. 2, pp. 182–194, 2014.

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 9163160, 14 pages
https://doi.org/10.1155/2018/9163160

Research Article

A Mobile-Based Question-Answering and Early Warning System for Assisting Diabetes Management

Wenxiu Xie,[1] **Ruoyao Ding**,[1] **Jun Yan**,[2] **and Yingying Qu**[3]

[1]*School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, China*
[2]*AI Lab, Yidu Cloud (Beijing) Technology Co. Ltd., Beijing, China*
[3]*School of Business, Guangdong University of Foreign Studies, Guangzhou, China*

Correspondence should be addressed to Yingying Qu; yingyinqu2@gmail.com

Received 27 March 2018; Accepted 26 April 2018; Published 6 June 2018

Academic Editor: Haoran Xie

With increasing demand for preventive management of chronic diseases in real time by using the Internet, interest in developing a convenient device on health management and monitoring has intensified. Unlike other chronic diseases, diabetes particularly type 2 is a lifelong chronic disease and usually requires daily health management by patients themselves. This study is to develop a mobile-based diabetes question-answering (Q&A) and early warning system named Dia-AID, assisting diabetes patients and populations at high risk. The Dia-AID system consists of three modules: a large-scale multilanguage diabetes frequently asked question repository, a multimode fusion Q&A framework, and a health data management module. A list of services including risk assessment and health early warning is provided to users for health condition monitoring. Using the diabetes frequently asked question repository as data, experiments are conducted on answer ranking and answer selection aspects. Results show that two essential methods in the system outperform baseline methods on both aspects.

1. Introduction

With the increasing attention of ubiquitous healthcare (U-healthcare) services and the developing of information technology, there has been a great need for preventive management of chronic diseases and management of individual health conditions [1]. Diabetes mellitus, a.k.a. diabetes, as one of the most representative chronic diseases, has become a serious global public health issue and the most challenging health problem in the 21st century [2–4]. The statistics of the number of diabetes patients 20-79 years of age in the past 18 years are shown in Figure 1, according to the latest global estimation from the International Diabetes Federation (IDF) and the Research2guidance report (http://www.research2guidance.com). Compared to 151 million in 2000, there is nearly a threefold increase in the number of adults living with diabetes mellitus. Moreover, the number is expected to increase from 425 million in 2017 to 629 million in 2045, which means that one out of 11 adults will suffer from diabetes [5]. In addition, as reported by the World Health Organization (WHO), diabetes was the direct cause of 1.6 million deaths in 2015. However, nearly 50% of diabetes patients are undiagnosed and remain unaware of their conditions. Among the patient population, the majority of diabetes cases are type 2 diabetes mellitus (T2DM) [6]. Unlike type 1 diabetes mellitus which remains unpreventable with current knowledge, 80% of type 2 diabetes mellitus can be prevented by keeping moderate blood sugar and lifestyle [7]. People with diabetes type 2 frequently need counseling on healthy diet and regular physical activity to reduce the risk of complication [8]. Thus, diabetes management is a crucial and necessary procedure for diabetes patients or people at a high diabetes risk [9–11].

Recently, the focus of healthcare is shifting from treatment to prevention and early diagnosis of disease [12]. Fox et al. [13] addressed that 31% of US smartphone owners used their phones to search for medical information online, 30% of Internet users consulted online reviews of rankings of healthcare services or treatments, and 26% of Internet users read other people's experiences about health or medical issues. By 2015, nearly 500 million smartphone users used mobile health applications especially for diet and disease

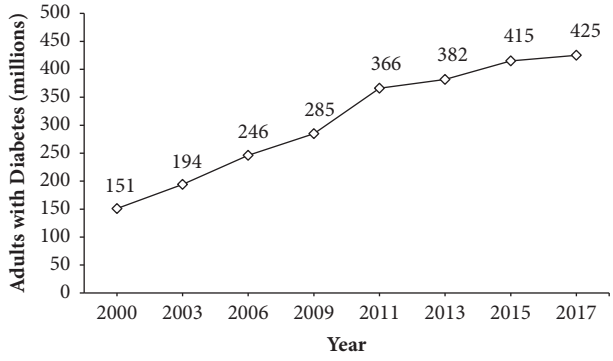


Figure 1: Total number of adults with diabetes (20-79 years old) around the world during 2000 to 2017.

management [7]. Later, Krebs et al. [14] showed that 58.23% (934/1604) of mobile phone users downloaded a health-related mobile app and used it at least once per day. As a convenient platform for checking users' health status on a real-time basis, mobile applications have been developed from information provision to lifestyle-oriented smart health management. Moreover, existing research presented that continuous real-time consulting and monitoring supported by smartphones is applicable for improving efficiency of diabetes self-management [7, 15–17]. Therefore, developing a mobile-based system for diabetes patients to assist in their health management is of great importance.

Many studies have shown that current medical search engines, e.g., PubMed, Medical Subject Headings (MeSH), and Unified Medical Language System (UMLS), are often unable to serve users with clinically relevant answers in a timely manner and thus fail to satisfy patients' counseling need [18, 19]. Hersh et al. [20] found that a healthcare professional took more than 30 minutes on average to seek an answer utilizing information retrieval systems. The process needs about 2 minutes on average to obtain an answer even for experienced doctors [21]. Instead, based on natural language processing techniques, question answering (Q&A) aims to provide users with direct, precise answers to their questions, and thus it is more preferred. Hence, there is an increasing demand to develop convenient and effective question-answering systems for the medical domain [21–24]. Moreover, there is a particularly growing demand of Q&A systems for effectively and efficiently assisting diabetes patients to better utilize ever-accumulating expert knowledge [1, 7, 15].

To that end, this study aims to develop a mobile-based question-answering and early warning system, called Dia-AID. The system consists of 3 modules: a large-scale multi-language diabetes FAQ repository, a multimode fusion Q&A framework, and a health data management module with early warning function. The repository captures diabetes questions with expert-defined answers and stores the question-answer knowledge in an interpretable and extendible form. The framework contains three different Q&A resolution strategies: knowledge-based Q&A, FAQ-based Q&A, and web-based Q&A. The health data management module containing early warning provides a convenient counseling service on a smart health platform to assist diabetes patients in monitoring their health conditions.

The contributions of this work include the following: (1) a large-scale multilanguage diabetes FAQ repository is built with a consistent representation format; (2) a novel multimode fusion Q&A framework that integrates three modes of Q&A technologies is proposed to fulfill diabetes information seeking need; (3) a health data management module containing early warning function is developed to monitor patient health status.

The rest of this paper is organized as follows. Section 2 introduces related work in biomedical question answering. Section 3 describes the mobile-based question-answering and early warning system Dia-AID in detail. Section 4 presents the experiment results of our methods based on the FAQ data repository. Section 5 addresses the conclusions.

2. Related Work

The aim of question answering is to provide precise answers instead of relevant documents from unstructured data sources to inquirers. The research of open domain question answering (Q&A) started from the prompt and instantiated work in the Text REtrieval Conference (TREC) evaluation campaign [25]. Recently, with the increasing demand of domain-specific applications, a growing interest has shifted from open domain Q&A to restricted domain Q&A [26, 27]. Molla et al. [28] addressed that restricted domain Q&A targeting domain-specific information was expected to achieve effective and reliable performance in real-world applications. Further, as claimed by Mishra et al. [29], restricted domain Q&A could fulfill the specialized information requirements of domain experts, therefore improving the satisfaction of users. Similarly, Yu et al. [30] and Rinaldi et al. [31] noted that restricted domain Q&A, such as biomedical domain [24, 32], could exploit domain-specific knowledge resources for deeper text analysis, as well as taking advantage of domain-specific typology formatting conventions to improve the answer extraction performance.

In light of Athenikos et al.'s research [27], medical domain question answering was facing the challenge of highly complex domain-specific terminology and lexical and ontological resources. Also emphasized by Abacha et al. [33], the key process was to translate the semantic relations expressed in questions into a machine-readable representation to analyze the natural language questions deeply and efficiently. They presented a complete question analysis approach including medical entity recognition, semantic relation extraction, and automatic translation to SPARQL queries. Result presented that 60% of the questions were correctly translated to SPARQL queries via the proposed method. Later, Anca [34] proposed the GFMed for dealing with the same problem and the challenge of querying a large number of Linked Data from various domains. GFMed was a Q&A system for biomedical interlinked data aiming to fill the gap between end users and formal languages by introducing a grammatical framework to translate biomedical information in natural language to the corresponding SPARQL language. The experimental results demonstrated that the proposed methodology for building

Controlled Nature Language for querying Linked Data was valid. Abacha et al. [35] proposed an approach for "Answer Search" based on semantic search and query relaxation to resolve the problem of automatic Q&A in medical domain. They defined question focuses as medical entities that were the most closely linked to answers to improve the overall performance of question answering. Terol et al. [36] claimed a general Q&A system that was capable of working over any restricted domain. Taking medical domain as an example application, their system answered medical questions according to a generic question taxonomy and gained 94.4% overall precision on the task.

During question-answering process, question representation is an essential step in question analysis and answer retrieval. Zhang et al. [37] proposed a system based on multilayer self-organizing map, providing an efficient solution to the organization problem of structured data of electronic books. A tree-structured representation was proposed to formulate the rich features of an e-book author. Their experiment results corroborated that the proposed models based on the tree-structured representation outperformed content-based models. Later in their further research, an efficient learning framework Tree2Vector for transforming tree-structured data into vectorial representations was proposed [38]. By utilizing Tree2Vector framework to map tree-structured book data into vectorial space, their continued experiments further presented that the mapped vectorial space could explore term spatial distributions over a book rather than the traditional document modeling methods [39].

A recent trend among medical Q&A systems was to incorporate the organized medical information throughout Q&A process in order to utilize the information for efficient health management in various areas such as U-healthcare [40, 41]. Jung et al. [42] developed a decision supporting method mainly for pain management for chronic disease patients based on frequent pattern tree mining. The proposed method aimed to reduce time and expenses for pain decision-making of patients who were frequently exposed to pain. Chung et al. [12] presented a knowledge-based health service by leveraging a hybrid wireless fidelity peer-to-peer architecture. The service was proposed to provide patients with efficient and economical healthcare through correct measurement of various biosignals, so that users could easily predict and manage both health and disease. Han et al. [43] introduced a U-health service system THE_MUSS, focusing on achieving reusability and resolvability, to provide stress and weight management services.

In subsequent medical Q&A developments, diabetes mellitus, as one of the top three major worldwide causes of death from noncommunicable diseases, has prompted numerous researches investigating the prevention, prevalence, and mortality of diabetes mellitus [15, 44–53]. There is a great demand for a Q&A system that can effectively and efficiently provide health consulting services and assist people in monitoring and managing their individual health conditions. Jung et al. [7] explored a mobile healthcare application for providing self-diabetes management to patients. By interoperating with Electronic Medical Record (EMR), the healthcare application provided services such as weight management, cardio-cerebrovascular risk evaluation, and exercise management. Waki et al. [54] developed a real-time interactive system DialBetics to achieve diabetes self-management, particularly HbA1c management. By an evaluation strategy, the system helped patients improve their HbA1c significantly by monitoring health data compared with continuing self-care regimen patients. More recently, Yoo et al. [1] proposed a Personal Health Record- (PHR-) based diabetes index service model through a mobile device, offering users a management information service for preventing diabetes. Users were able to check their health conditions on a real-time basis and receive information about desirable health behaviors and dietary habits related to diabetes.

Yet, the existing diabetes management applications provided general information search and management, while ignoring counseling services, which were crucial for managing the health condition of diabetes patients. Besides, as claimed by Mishra et al. [29], the cons of restricted domain Q&A included the limited repository of domain-specific questions. To overcome these difficulties, we built a LMD-FAQ repository to provide users with concise and accurate answers by physicians or experts from debates related professional websites. Moreover, we aim to leverage the LMD-FAQ repository to provide counseling services of diet, medication, and symptoms for diabetes patients. In addition, based on our previous work [55], by analyzing global clinical trials of 190 countries provided by the National Institutes of Health (NIH), we discovered 6 representative health characteristics that were closely related to diabetes mellitus to better manage users' health conditions. The six representative health characteristics were Body Mass Index (BMI), Glucose, Systolic Hypertension, Diastolic Hypertension, HbA1c, and creatinine. Further, we defined several early warning intervals of health characteristics by referring to the existing international medical standards for health management and risk warning.

3. Methods and Materials

The architecture of our mobile-based diabetes question-answering and early warning system Dia-AID is shown in Figure 2. It consists of 3 modules: a large-scale multilanguage diabetes FAQ repository (LMD-FAQ repository), a novel multimode fusion question-answering framework (MMF-QA), and a diabetes data management module with early warning (DM-EW). The LMD-FAQ repository contains a large number of diabetes question-answer pairs acquired from mainstream diabetes-related professional websites. The MMF-QA framework integrates three strategies: knowledge-based Q&A, FAQ-based Q&A, and web-based Q&A. The DM-EW module records patients' health data and monitors their health conditions in real time. Six representative health characteristics that are closely related to diabetes mellitus, that is, BMI, glucose, systolic hypertension, diastolic hypertension, HbA1c, and creatinine, are applied. In case of a rapid characteristic change or a predication of deterioration, the module will automatically warn patients and provide them with dietary guidelines.

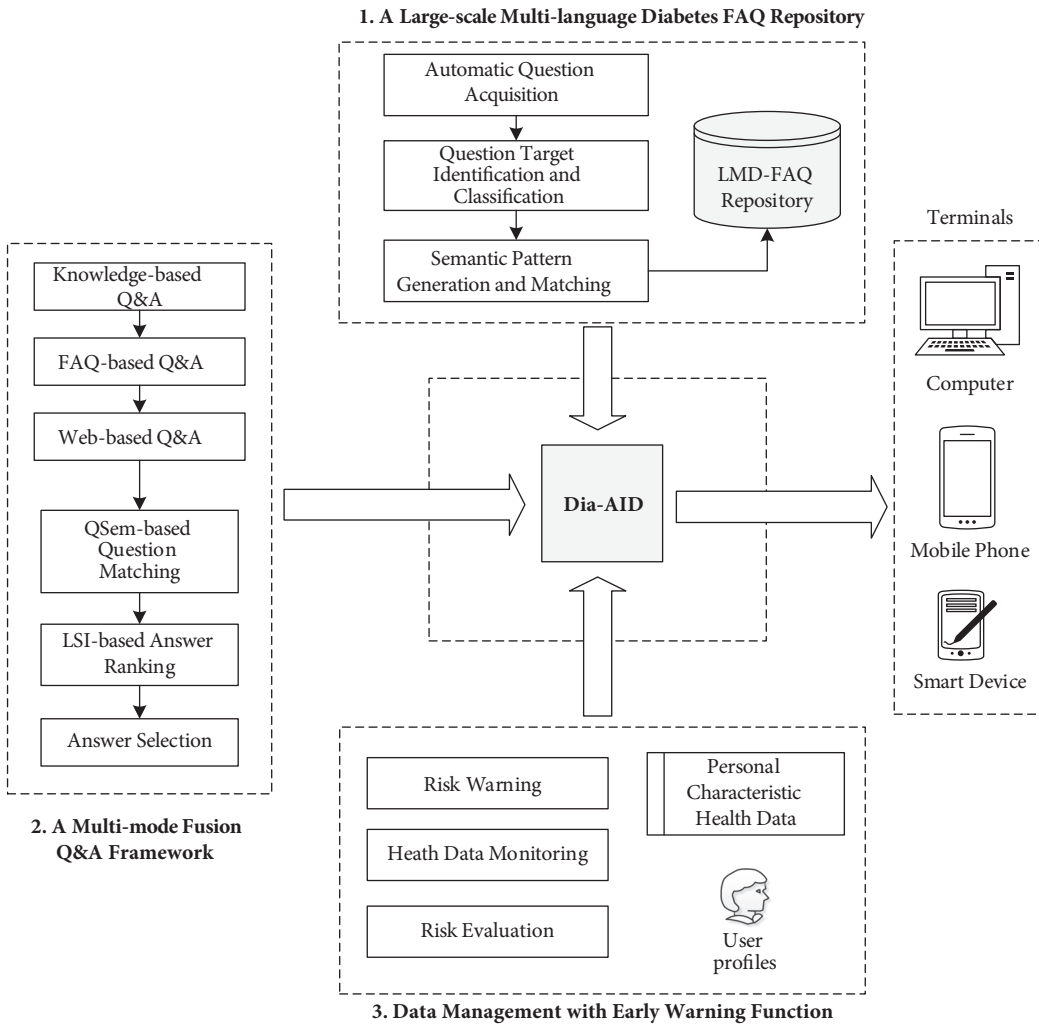


Figure 2: The architecture of the diabetes question-answering and early warning system, Dia-AID.

3.1. The Large-Scale Multilanguage Diabetes FAQ Repository. Frequently Asked Questions (FAQs) provide specific answers to the questions that are frequently asked when users browse specific websites. For example, the website Health China (http://health.china.com) allows users to ask questions in free text and those who are experts in the field answer the questions freely. These questions with professional answers are collected and organized as FAQ data. The FAQ data can dramatically benefit question answering by reusing the accumulated professional knowledge. In this paper, we develop a method to automatically construct a large-scale multilanguage diabetes FAQ (LMD-FAQ) repository through identifying FAQ data from professional diabetes websites.

As illustrated in Figure 2, our method includes four steps: (1) The first step is automatic question acquisition. We first analyze the page structures of specific websites to identify diabetes questions. The websites, elaborately selected by domain experts, include Diabetes Clinical Guidelines (Chinese Medical Association Diabetes Branch), professional diabetes websites (International Diabetes Union, the American Diabetes Association, etc.), diabetes professional information websites (CDC Health Channel), and diabetes interactive question-answering websites (Yahoo! Knowledge). The questions and associated answers are then extracted using regular expression matching with the page codes. (2) The second step is question target identification and classification. Based on our previous work [56, 57], an automated answer type identification and classification method is applied to extract the target and intent of questions by utilizing both syntactic and semantic analysis. Considering that syntactic structures vary according to the ways questions are asked, four typical situations are identified and analyzed with each of them having a specific processing strategy. During the process, question target features are extracted via a principle-based syntactic parser and then expanded with their hypernymy features and semantic labels. Finally, the expanded features are sent to a trained classifier to predict corresponding answer types. (3) The third step is semantic pattern generation and matching. Semantic pattern is utilized to index the questions with answers in a more structured and semantic way. With question target and answer type extracted by the second step, the questions are represented by a structural semantic

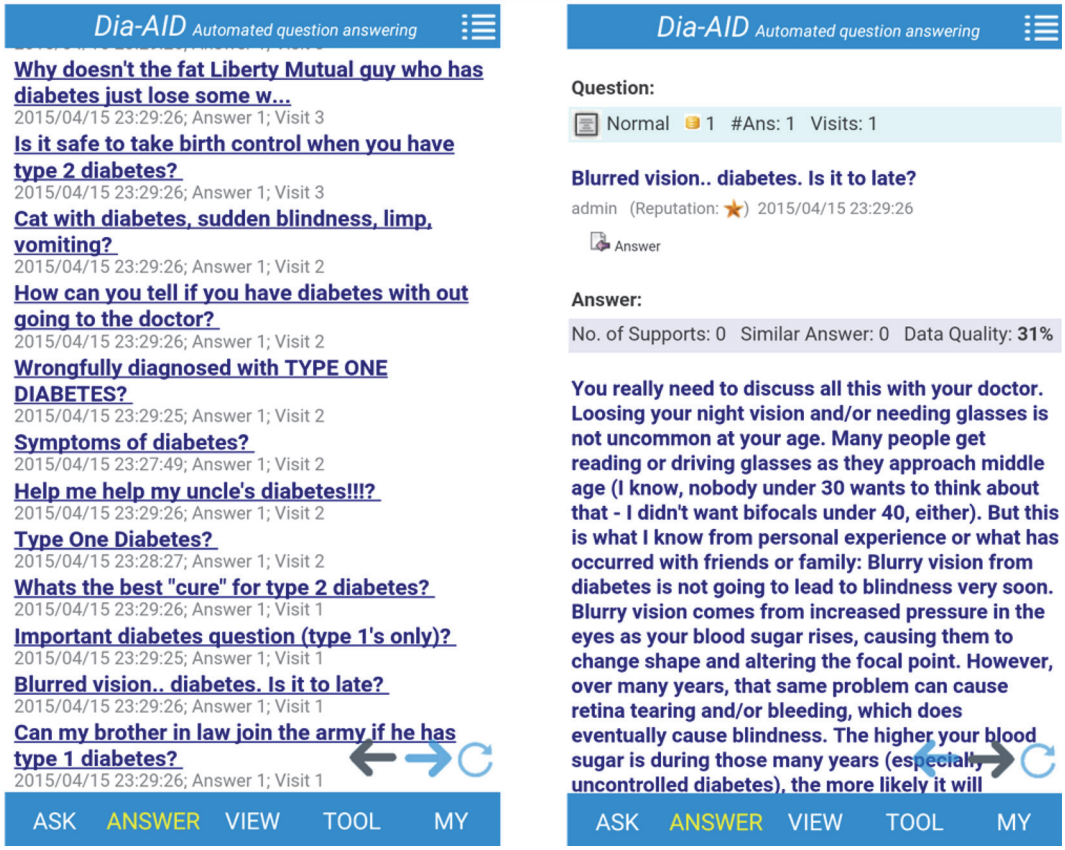


Figure 3: The visualization of example FAQ data in the LMD-FAQ repository.

pattern, which consists of five components: the question target, question type, concept, event, and constraint. An entropy-based method proposed in previous work [58] is applied for automated semantic pattern generation. Figure 3 shows the visualization of example FAQ data in the LMD-FAQ repository.

Based on the above procedure, the method extracts FAQ data from professional websites, formats them using a consistent representation, and indexes them with semantic patterns for fast retrieval. Through the automatic process and the human review on the indexed data, the FAQ repository can be incrementally maintained. Currently, the LMD-FAQ repository comprises 19,317 English frequently asked QA pairs and 6,041 Chinese QA pairs. The repository provides our Q&A system with fundamental data support for answering commonly posted questions.

3.2. The Multimode Fusion Question-Answering Framework. The multimode fusion question-answering framework (MMF-QA) integrates three Q&A models: knowledge-based Q&A, FAQ-based Q&A, and web-based Q&A. The overall framework is shown in Figure 4. The procedure of the models is described as follows.

The knowledge-based Q&A model relies on a diabetes knowledge base to generate concise answers for posted questions. For a new given question, the model analyzes the structure and keywords of the question and then generates a corresponding semantic pattern. Thus, the question is transformed from natural language to a structural semantic representation that captures semantic information such as question target, question type, concept, event, and constraint The question then is further represented as a tuple: $([Concept_1], \{Relation\}, [Concept_2])$, in which "$Concept_1$" and "$Concept_2$" are used to label meaningful entities. The represented question is used for answer extraction from knowledge base. For instance, "*What's the symptoms of diabetes?*" is represented as ([*symptoms*], {*Rel: of*}, [*diabetes*]). Therefore, the knowledge-based Q&A process mainly maps entities and their relations to formally represented tuples, which are further used to match knowledge base to retrieve accurately matched knowledge elements as answers.

The FAQ-based Q&A model computes matching scores between a given question and questions in the FAQ repository. The questions with matching scores larger than a specific threshold are kept as candidates. The candidate questions then are ranked and the top k questions with the highest scores are returned. The model consists of three main steps: Qsem-based question matching, LSI-based answer ranking, and answer selection. As claimed by [59], a major challenge of FAQ-based Q&A is to match questions to corresponding question-answer pairs. Here, we apply a QSem-based question matching framework, proposed in one of our previous works [60], to support answering FAQs through reusing accumulated QA data. The framework considers

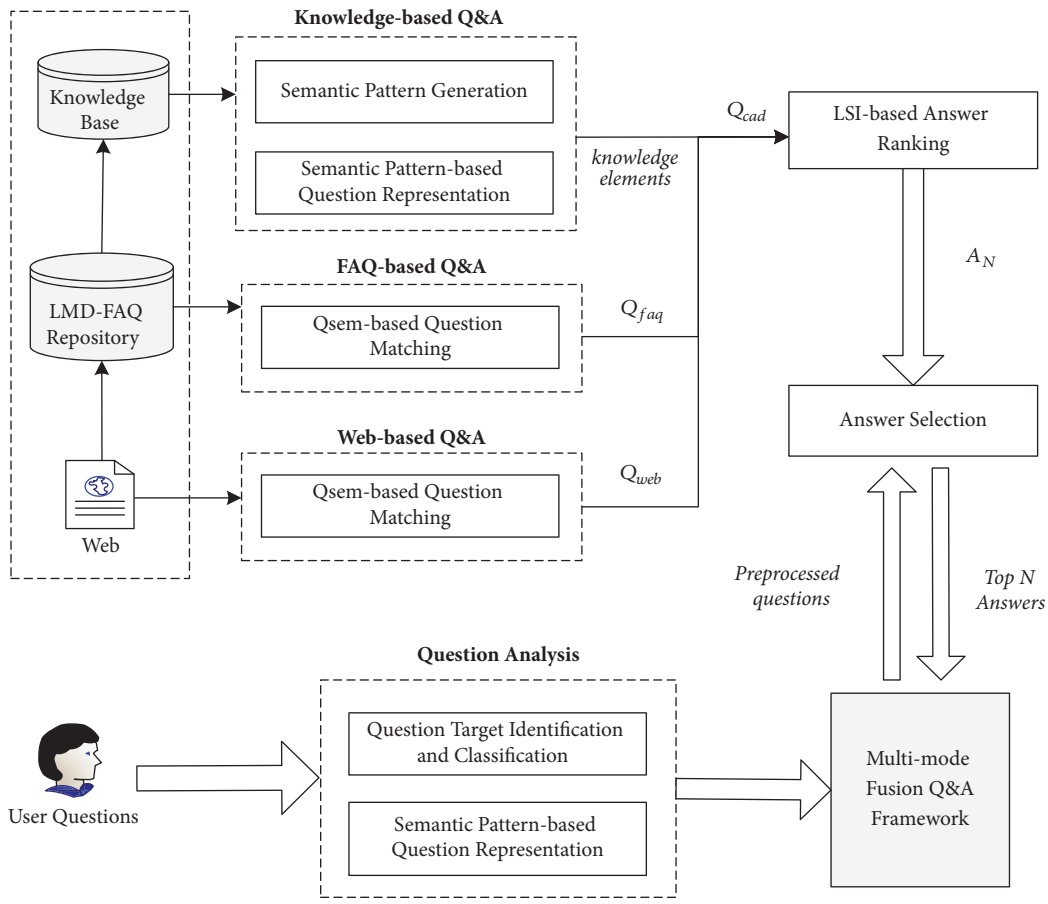


Figure 4: The multimode fusion Q&A framework, MMF-QA.

both question word types and semantic pattern according to their functionalities in question matching. The question word types include question target word, user-oriented word, and irrelevant word. These three word types are semantically labeled by a predefined ontology to enrich the semantic representation of questions. For each word type, different similarity strategies are applied to calculate the similarity, as described in [60]. The similarity calculations for question target and user-oriented word type between question q_i and a FAQ candidate faq_j are shown in (1), (2), and (3), respectively.

$$Simi_{QT}\left(q_i, faq_j\right) = \frac{2 \times \left|q_i\left(QTW \longrightarrow L^*\right) \cap faq_j\left(QTW \longrightarrow L^*\right)\right|}{\left|q_i\left(QTW \longrightarrow L^*\right)\right| + \left|faq_j\left(QTW \longrightarrow L^*\right)\right|} \quad (1)$$

$$SMatch\left(w_m, w_n\right) = \begin{cases} 0, & \left|\left(w_m \cup w_m \longrightarrow S\left(w_m\right)\right) \cap \left(w_n \cup w_n \longrightarrow S\left(w_n\right)\right)\right| = 0 \\ 1, & \left|\left(w_m \cup w_m \longrightarrow S\left(w_m\right)\right) \cap \left(w_n \cup w_n \longrightarrow S\left(w_n\right)\right)\right| \geq 1 \end{cases} \quad (2)$$

$$Simi_{UO}\left(q_i, faq_j\right) = \frac{2 \times \sum SMatch\left(q_i\left(UOW\right), faq_j\left(UOW\right)\right)}{\left|q_i\left(UOW\right)\right| + \left|faq_j\left(UOW\right)\right|} \quad (3)$$

In the equations, $Simi_{QT}$ denotes the similarity score of QT word type between a given new question q_i and an existing FAQ question faq_j. L^* denotes the set of semantic labels corresponding to target words of the question. $q_i(QTW \rightarrow L^*)$ and $faq_j(QTW \rightarrow L^*)$ represent the semantic labels of QT words in q_i and faq_j through semantic labeling, respectively. $w_m \cup w_m \rightarrow S(w_m)$ denotes synonymy words expansion of word w_m. *SMatch* denotes the synonymy-based word matching of two words w_m and w_n. $w_m \cup w_m \rightarrow S(w_m)$ is the synonymy extension of word w_m by adding synonymy word collection $S(w_m)$. By integrating the previous three parts of matching, the overall matching score $Simi_{SC}(q_i, faq_j)$ of the two questions q_i and faq_j through balancing the similarity of each part is calculated as shown in

$$MatchSC\left(q_i, faq_j\right) = \alpha \times Simi_{QT} + \beta \times Simi_{UO} + \left(1 - \alpha - \beta\right) \times Simi_{SP} \quad (4)$$

After question matching, top k FAQs with the highest matching scores are selected as candidates set Q_{faq}. Meanwhile, the web-based Q&A model uses a similar strategy to compute the matching scores to web question collections. It extracts k answers from websites via the standard question-answering techniques. Similarly, the web-based Q&A returns a candidate question-answer set Q_{web}. Q_{web} and Q_{faq} are

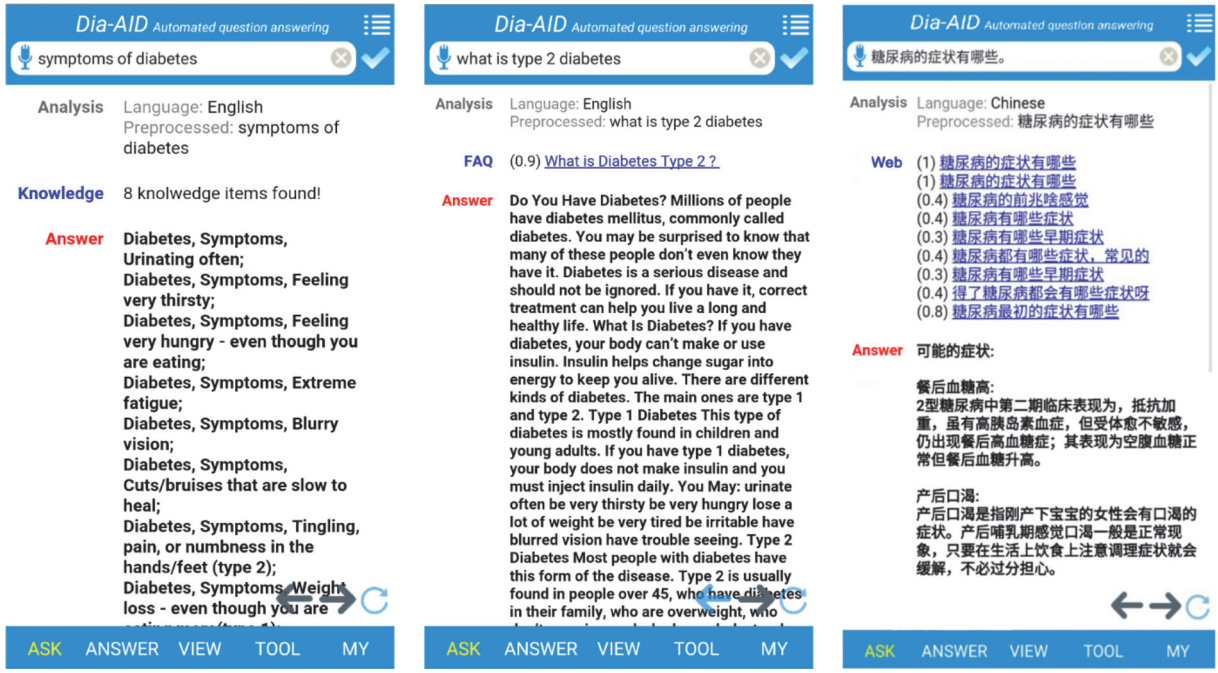


FIGURE 5: The screen snapshots of the mobile-based system for providing diabetes information services using the three Q&A models.

merged as the final answer candidates Q_{cad} for answer ranking and answer selection.

We propose a LSI-based answer ranking method to re-rank the questions in Q_{cad}. The ranking method consists of three steps: feature extraction, Latent Semantic Indexing (LSI) similarity calculation, and ranking. The extracted features of Chinese questions are bag-of-words (BOWs) and Character, while the features for English questions are bag-of-words feature only. The LSI approach takes advantage of implicit higher-order semantic structure and matches words in queries with words in documents [61]. Here we treat each candidate answer as a short document and detect the most relevant answers via the LSI-based method. After that, the candidate answers are reranked based on the similarity values and the top N answers as candidate list A_N are returned.

Finally, there is an answer selection process. The selection of a candidate answer as correct or incorrect can be treated as a binary classification task. The question and corresponding top N candidate answers in list A_N are transformed to N QA pairs. We propose an answer selection approach via a Logistic Regression (LR) classifier, which includes four steps: feature extraction, parameter tuning, model training, and answer selecting. Using the features similar to the LSI-based approach, QA pairs are randomly selected from the LMD-FAQ repository as training data. The QA pairs with correct answers are labeled as "1", and "0" otherwise. We then tune the parameter "C" (inverse of regularization strength) to avoid overfitting/underfitting issue. After parameter optimization, the best parameter is applied in the LR classifier, which then is applied to select the best N candidate answers, where the top 1 is the best answer and the remaining N-1 answers in list A_N are relevant answers. Figure 5 shows the screen snapshots of the knowledge-based Q&A, FAQ-based Q&A, and web-based Q&A modes.

3.3. Diabetes Data Management with Early Warning. Since diabetes patients and people at high risk usually need long-term health management, we develop a real-time data management module incorporating early warning to achieve patient health self-management.

In the data management module, users are required to register their basic information. After that, the users can log in to report their recent health data related to six main characteristics: HbA1c, BMI, glucose, systolic hypertension (hypertension_S), diastolic hypertension (hypertension_D), and creatinine. The health data then are stored in server side securely.

With the historical health data, the module calculates and monitors the health status in real time. For each of the characteristics, we set an alarm value according to literature review on IDF documents and reports. Once the health data has a dramatic change or the characteristics are close to their corresponding alarm value ranges, the system will automatically deliver a warning message to the users about the situation. To evaluate the usability of the system, a 2-month randomized study is designed. Thirty people volunteered as internal test users to monitor their health condition via the Dia-AID system. During the test, users measure and report the data of the six characteristics by themselves. Based on each new data report, the system calculates the existing data and newly submitted data to make a summarization of the health condition in real time. Table 1 shows the reported health data records by a user *Cecil*.

The system records all the reported health data and generates data change curves automatically. For example, Figure 6

Table 1: The reported health data records of the user *Cecil*.

Characteristic	Value	Time	Characteristic	Value	Time
HBA1C	7.5	2017-11-01 10:22:03	HBA1C	9.0	2017-11-04 19:15:18
GLUCOSE	12.0	2017-11-01 10:22:03	BMI	22.0	2017-11-04 19:15:18
BMI	22.1	2017-11-01 10:22:03	GLUCOSE	15.0	2017-11-04 19:15:18
HYPERTENSION_S	111.0	2017-11-01 10:22:03	HYPERTENSION_S	113.0	2017-11-04 19:15:18
HYPERTENSION_D	82.0	2017-11-01 10:22:03	HYPERTENSION_D	84.0	2017-11-04 19:15:18
CREATININE	1.18	2017-11-01 10:22:03	CREATININE	1.24	2017-11-04 19:15:18
HBA1C	6.0	2017-11-02 18:24:39	HBA1C	9.0	2017-11-05 18:10:25
BMI	22.0	2017-11-02 18:24:39	BMI	24.0	2017-11-05 18:10:25
GLUCOSE	10.0	2017-11-02 18:24:39	GLUCOSE	15.0	2017-11-05 18:10:25
HYPERTENSION_S	110.0	2017-11-02 18:24:39	HYPERTENSION_S	113.0	2017-11-05 18:10:25
HYPERTENSION_D	80.0	2017-11-02 18:24:39	HYPERTENSION_D	84.0	2017-11-05 18:10:25
CREATININE	1.2	2017-11-02 18:24:39	CREATININE	1.24	2017-11-05 18:10:25
HBA1C	7.8	2017-11-03 18:30:28	HBA1C	10.0	2017-11-06 19:21:13
BMI	21.9	2017-11-03 18:30:28	BMI	24.0	2017-11-06 19:21:13
GLUCOSE	12.0	2017-11-03 18:30:28	GLUCOSE	18.0	2017-11-06 19:21:13
HYPERTENSION_S	111.0	2017-11-03 18:30:28	HYPERTENSION_S	115.0	2017-11-06 19:21:13
HYPERTENSION_D	82.0	2017-11-03 18:30:28	HYPERTENSION_D	86.0	2017-11-06 19:21:13
CREATININE	1.2	2017-11-03 18:30:28	CREATININE	1.15	2017-11-06 19:21:13

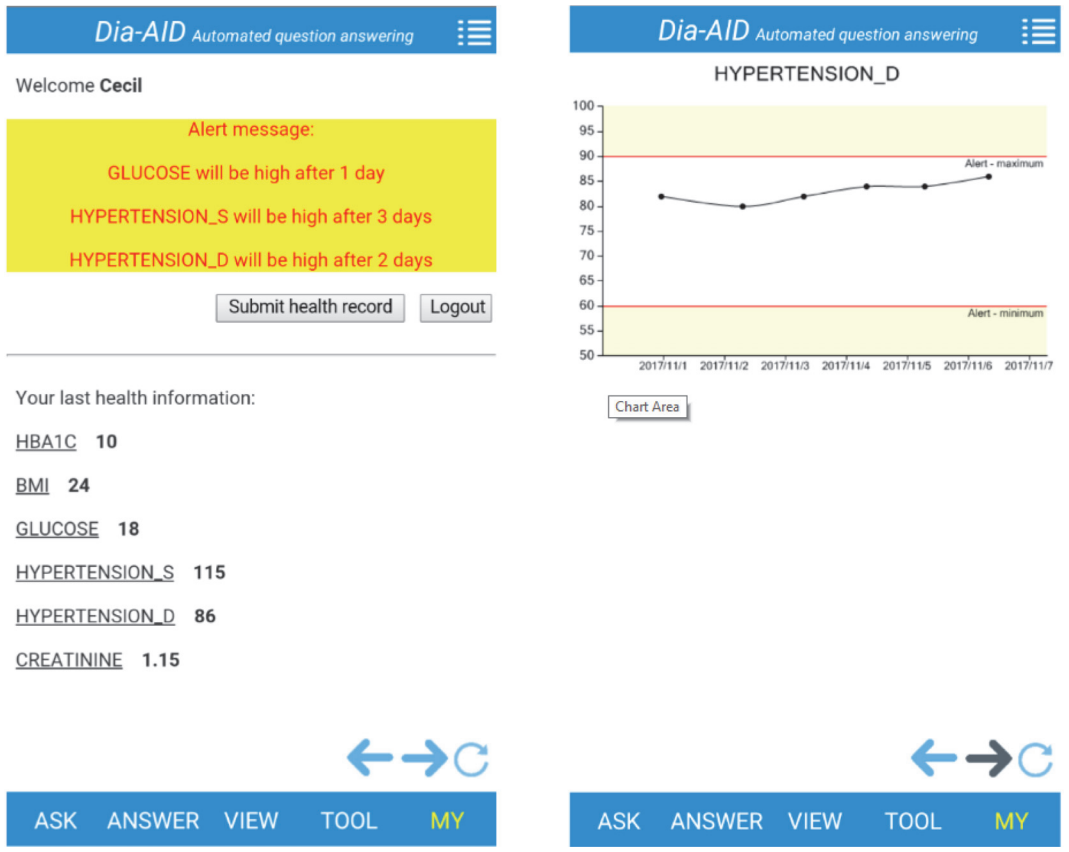


Figure 6: The health data management with early warning function and health record visualization by time for the user *Cecil*.

shows the trend curve of *Cecil*'s diastolic hypertension in the last 7 days. When the current newly submitted health data is within safe range and there is no dramatic change compared with last report, the system shows the user with the health status messages, e.g., "Your health status is good" in green color. Once the system identifies current user data exceeding alarm range (either too high or too low) according to the current change trend, the system will evaluate how long it takes to reach the alarm value. The system will evaluate how long it will take to reach the alarm value. If the period is too short, the system will automatically warn the current user. For example, the system warns the user *Cecil* that diastolic hypertension is too high and will be in a danger range after 2 days if the user does not have any control on it. Through the health data management incorporating early warning, users can review their health status and take actions to reduce the risk of diabetes according to the warning messages.

4. Results

4.1. Datasets. Since there is no available diabetes FAQ dataset for evaluation, the evaluations of the proposed LSI-based answer ranking approach and answer selection method were based on the constructed LMD-FAQ repository. To test the LSI-based answer ranking approach, we randomly selected 500, 750,1000,1250,1500, and 1750 Chinese question-answer tuples (*question,* <*answer-set*>) from the repository, respectively, as six subdatasets of **Evaluation dataset**-**A**. For each question-answer tuple, it contains one question and an answer set which consists of one correct answer and nine incorrect answers randomly generated from the rest of the repository. Thus, each question contains 10 candidate answers for ranking. For answer selection evaluation, we suppose each question has k candidate answers; i.e., for each question, *k-1* incorrect answers are randomly generated as negative samples. In this paper, k is set to 5 and 10. For the setting k=5, 6000 QA pairs are randomly generated as **Training dataset**-**B1**, and 2500 QA pairs are randomly generated as **Testing dataset**-**C1**. For the setting k=10, 8000 QA pairs are randomly generated as **Training dataset**-**B2**, and 5000 QA pairs are randomly generated as **Testing dataset**-**C2**.

4.2. Evaluation Metrics. The evaluation metrics include Mean Reciprocal Rank (MRR), Accuracy@N of the returned answers, precision, recall, and F1 measure, all of which are commonly used metrics to evaluate the performance of Q&A systems.

(i) MRR: Mean Reciprocal Rank of the first correct answer, as shown in (5) (i.e., 1 if a correct answer was retrieved at rank 1, 0.5 if a correct answer was retrieved at rank 2, and so on. Q is the test set and $|Q|$ denotes the number of questions in Q. $rank_i$ represents the position of the first correct answer in answer ranking candidates to a test question Q_i).

(ii) Accuracy@N: proportion of correct answers in top N returned answers by the system, as shown in (6) ($C_i(N) = 1$ if there is at least one correct answer in top N candidates; otherwise, it is 0).

(iii) Precision for any of the categories is the number of true positives (TP) (i.e., the number of questions correctly labeled as belonging to the positive categories) that are divided by the total number of questions labeled as belonging to the positive categories, as shown in (7). False positive (FP) is the number of questions that the system incorrectly labeled.

(iv) Recall is defined as the number of true positives divided by the total number of questions that actually belong to positive categories (i.e., the sum of true positive and false negative), as shown in (8).

(v) F1-measure considers both the precision and the recall to compute a balanced score, as shown in (9).

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \tag{5}$$

$$Accuracy@N = \frac{1}{|Q|}\sum_{i=1}^{|Q|}C_i(N) \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

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

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{9}$$

4.3. Results. To validate the proposed LSI-based answer ranking method, we conduct the following two experiments. The first experiment is to verify the effectiveness of the LSI-based answer ranking method by comparing to five baselines. We adopt Doc2Vec, Latent Dirichlet Allocation (LDA), Locality Sensitive Hashing (LSH), docsim, and Synonyms [62] as baselines. We randomly select 500 questions and measure the performance in MRR and Accuracy@N (Acc@N, $N = 1, 2, 3, 4, 5$). Compared with the baselines, our method achieves the best performance in all evaluation metrics, as shown in Table 2. For MRR, our method improves by 17.80% compared to LSH which has the best performance among baselines. For Acc@1, LSH also obtains the best performance as 0.6733 among baselines. Our method outperformed LSH with an improvement of 23.52%. In addition, our method ranks 94.99% of the correct answers in the top five of candidate answers. The improvements of MRR and Acc@1 prove that the proposed method can potentially promote the positions of correct answers.

To assess the stability of the proposed method, the second experiment is conducted by comparing to the same five baselines with the measures of MRR and Acc@1. The used dataset is Evaluation dataset-A. Figure 7 illustrates the experiment results measured in MRR, while Figure 8 shows the results measured in Acc@1. From the result, our method achieves stable performance over all different sizes of the question sets. This result is promising since our method ranks most of the correct answers in the top of the candidate answer list. Moreover, compared to the baselines, our method gains

Table 2: Evaluation results compared to baselines.

	MRR	Acc@1	Acc@2	Acc@3	Acc@4	Acc@5
Doc2Vec	0.2811	0.0721	0.1923	0.2905	0.3967	0.5250
LDA	0.5536	0.3326	0.5811	0.7294	0.8016	0.8597
Synonyms	0.6082	0.4108	0.6332	0.7454	0.8376	0.8877
docsim	0.7233	0.6312	0.7074	0.7515	0.7895	0.8336
LSH	0.7517	0.6733	0.7394	0.7835	0.8056	0.8276
Our Method	**0.8855**	**0.8317**	**0.8918**	**0.9259**	**0.9459**	**0.9499**

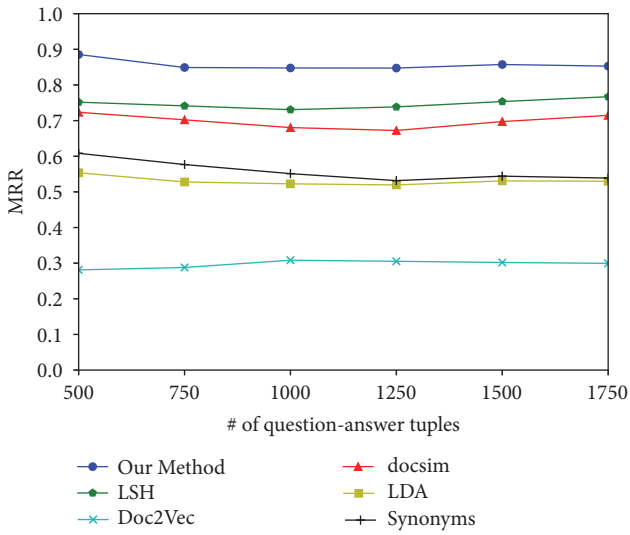


Figure 7: The performance comparison between the proposed method and baselines with the increasing number of question-answer tuples using the MRR measure.

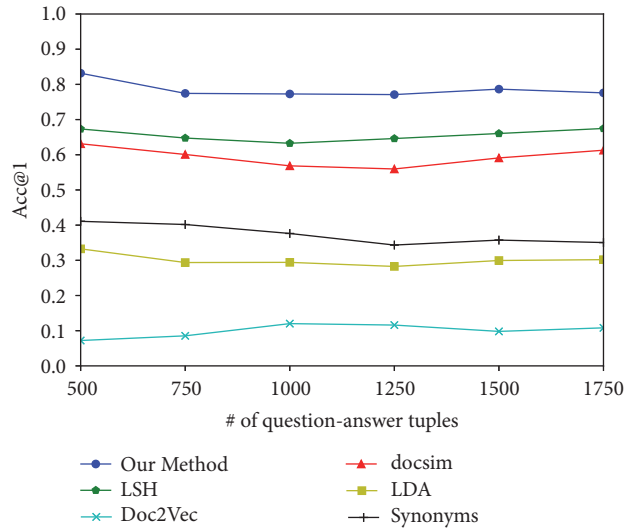


Figure 8: The performance comparison between the proposed method and baselines with the increasing number of question-answer tuples using the Acc@1 measure.

the best performance measured in Acc@1 on all the question sets. From the results, even with the increasing number of questions, nearly 85% of correct answers are ranked in the top of the candidate answer list.

Since our answer selection approach uses a binary classifier, we assess the method by evaluating the effectiveness of answer classification. During the evaluation, three experiments are designed: the first is to train optimized parameters, the second aims to assess the stability of classification, and the third aims to evaluate the effectiveness by comparing with baseline methods. The datasets used for evaluation are from the constructed LMD-FAQ repository and the evaluation metrics are precision, recall, F1, and accuracy.

To avoid the overfitting/underfitting problem, we tune the parameter "C" (inverse of regularization strength) for the LR classifier as described above. 12,651 QA pairs are randomly selected from the LMD-FAQ repository as the dataset. The dataset then is randomly shuffled into two subgroups as training (70%) and testing (30%). We use k-fold cross-validation to assess the model performance. Figure 9 demonstrates the validation curve, where training accuracy represents the results on testing dataset and validation accuracy denotes the 10-fold cross-validation results. From the results, the method gains the best performance when "C" is equal to 1, which is the best parameter applied in the following two experiments.

The stability of the proposed method is tested with different sizes of training data and different k values. By setting k=5, the Training dataset-B1 is randomly divided into 5 training subsets containing 2000, 3000, 4000, 5000, and 6000 question-answer pairs, respectively. Similarly, by setting k=10, the Training dataset B2 is randomly divided into 5 training subsets with 4000, 5000, 6000, 7000, and 8000 question-answer pairs. The datasets C1 and C2 are used as testing datasets independently. The results are measured in accuracy (Acc), precision, recall, and F1-measure (F1). As shown in Figure 10, our method receives a stable performance on all evaluation metrics with k=5. When the size of training dataset is larger than 3000, the performance on all metrics increases. The experiment results indicate that our method is not affected much by training dataset size. As illustrated in Figure 11, the performance measured in accuracy remains stable on all sizes of training datasets. With the increasing of training dataset size, the performance measured in F1 increases. Comparing the performance on the two dataset settings, our method yields a better performance when k equals 10, which indicates that the proposed method remains stable even with more incorrect answers in candidate answer lists.

We further compare our method with five commonly used classification methods: Support Vector Machine (SVM),

Table 3: The comparison of our answer selection method with baseline methods using different k settings.

Setting	Methods	Accuracy	Precision	Recall	F1
k=5	KNN	0.8052	0.6820	0.5679	0.5746
	GaussianNB	0.7383	0.6980	0.8026	0.6956
	RF	0.8301	0.7339	0.7097	0.7203
	SVM	0.8373	0.7513	0.8029	0.7706
	PPN	0.8842	0.8129	0.8540	0.8305
	Our method	**0.9222**	**0.8859**	**0.8657**	**0.8753**
k=10	KNN	0.9032	0.8076	0.5258	0.5247
	RF	0.9106	0.7523	0.7277	0.7391
	GaussianNB	0.8956	0.7197	0.7737	0.7422
	SVM	0.9154	0.7642	0.8399	0.7951
	PPN	0.9271	0.7897	**0.8633**	0.8206
	Our method	**0.9651**	**0.9415**	0.8559	**0.8929**

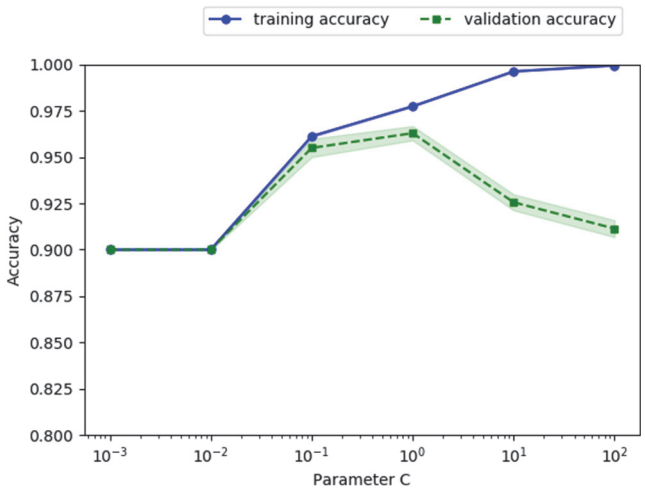


Figure 9: The validation curve of parameter "C" using different values.

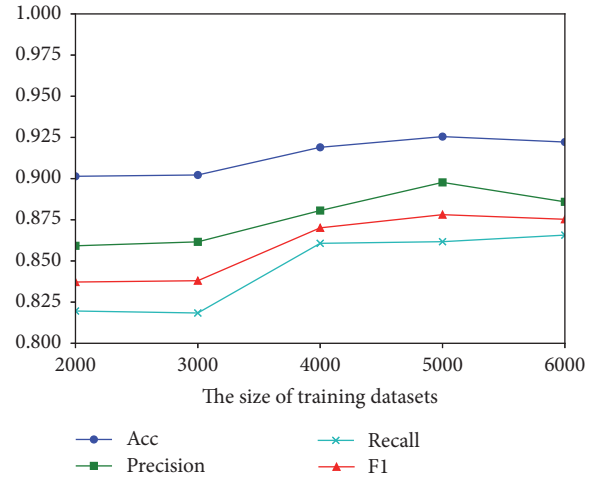


Figure 10: The performance with the increasing size of training datasets when k=5.

Perceptron (PPN), Random Forest (RF), Gaussian Naive Bayes (GaussianNB), and k-Nearest Neighbor (KNN). The datasets used are the Training dataset-B1 and Training dataset-B2 and the corresponding Testing dataset-C1 and Testing dataset-C2. The evaluation metrics are accuracy, precision, recall, and F1. Table 3 shows the comparison results using different dataset settings. By setting k=5, an accuracy of 0.9222, a precision of 0.8859, a recall of 0.8657, and an F1 of 0.8753 are achieved as the best performance compared to five baseline methods. By setting k=10, our method also obtains the highest performance on all evaluation metrics compared to the baselines. Particularly, the higher precision and F1 are more preferable since our expectation is the return of more correct answers to users to improve user satisfaction.

5. Conclusions

Aimed at assisting diabetes patients or populations at high risk of diabetes to have long-term health management, this paper designed and developed a mobile-based question-answering and early warning system, Dia-AID. The system

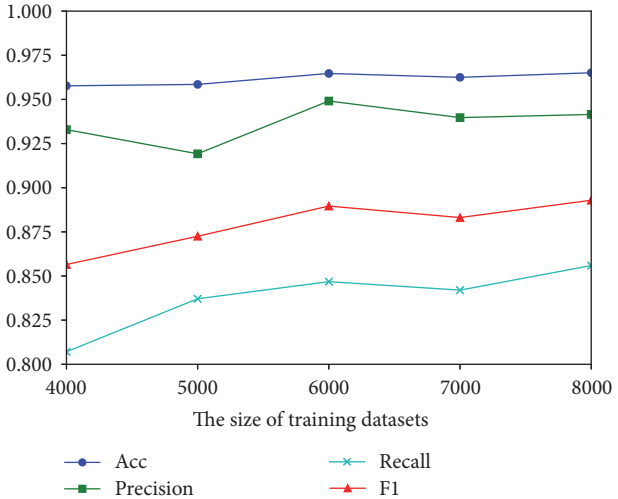


Figure 11: The performance with the increasing size of training datasets when k=10.

assists users in providing diabetes information and monitoring their health status through diabetes question answering, risk assessment, and health record management. We evaluated two essential models in our system and compared them with five baseline methods on various metrics. The results showed that our methods achieved the best performance compared with the baseline methods.

Data Availability

The diabetes data is not made publicly available.

Conflicts of Interest

There are no conflicts of interest in this paper.

Acknowledgments

The work was substantially supported by a grant from the National Natural Science Foundation of China (no. 61772146), the Science and Technology Plan of Guangzhou (no. 201804010296), the Innovative School Project in Higher Education of Guangdong Province (no. YQ2015062), and Scientific Research Innovation Team in Department of Education of Guangdong Province (no. 2017KCXTD013).

References

[1] H. Yoo and K. Chung, "PHR Based Diabetes Index Service Model Using Life Behavior Analysis," *Wireless Personal Communications*, vol. 93, no. 1, pp. 161–174, 2017.

[2] P. Zimmet, K. G. M. M. Alberti, and J. Shaw, "Global and societal implications of the diabetes epidemic," *Nature*, vol. 414, no. 6865, pp. 782–787, 2001.

[3] F. Aguiree and A. Brown, "IDF Diabetes Atlas," in *IDF Diabetes Atlas - sixth Edition*, International Diabetes Federation, Belgium, 2013.

[4] P. Z. Zimmet, D. J. Magliano, W. H. Herman, and J. E. Shaw, "Diabetes: a 21st century challenge," *The Lancet Diabetes & Endocrinology*, vol. 2, no. 1, pp. 56–64, 2014.

[5] J. G. Melton, *IDF Diabetes Atlas*, 8th edition, 2017.

[6] G. Roglic, "WHO Global report on diabetes: a summary," *International Journal of Noncommunicable Diseases*, vol. 1, no. 1, pp. 3–8, 2016.

[7] E.-Y. Jung, J. Kim, K.-Y. Chung, and D. K. Park, "Mobile healthcare application with EMR interoperability for diabetes patients," *Cluster Computing*, vol. 17, no. 3, pp. 871–880, 2014.

[8] World Health Organization, "Global Report on Diabetes," *Isbn*, vol. 978, article 88, 2016.

[9] J. Beck, D. A. Greenwood, L. Blanton et al., "2017 National Standards for Diabetes Self-Management Education and Support," *Diabetes Care*, article dci170025, 2017.

[10] A. D. American Diabetes Association, "4. Lifestyle Management," *Diabetes Care*, vol. 40, no. Suppl. 1, pp. S33–S43, 2017.

[11] Diabetes Prevention Program Research Group, "10-year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study," *The Lancet*, vol. 374, no. 9702, pp. 1677–1686, 2009.

[12] K. Chung, J.-C. Kim, and R. C. Park, "Knowledge-based health service considering user convenience using hybrid Wi-Fi P2P," *Information Technology and Management*, vol. 17, no. 1, pp. 67–80, 2016.

[13] Susannah Fox and Maeve Duggan, "Health Online 2013 — Pew Research Center," *National survey by the Pew Research Center's Internet and American Life Project*, 2013. http://www.pewinternet.org/2013/01/15/health-online-2013/.

[14] P. Krebs and D. T. Duncan, "Health app use among US mobile phone owners: a national survey," *JMIR mHealth and uHealth*, vol. 3, no. 4, article e101, 2015.

[15] S. Chavez, D. Fedele, Y. Guo et al., "Mobile apps for the management of diabetes," *Diabetes Care*, vol. 40, no. 10, pp. e145–e146, 2017.

[16] K. Waki, H. Fujita, Y. Uchimura et al., "DialBetics: Smartphone-based self-management for type 2 diabetes patients," *Journal of Diabetes Science and Technology*, vol. 6, no. 4, pp. 983–985, 2012.

[17] P. P. Committee and A. Classification, "Standards of medical care in diabetes—2010," *Diabetes Care*, vol. 33, no. S1, pp. S11–S61, 2010.

[18] D. Demner-Fushman and J. Lin, "Answer extraction, semantic clustering, and extractive summarization for clinical question answering," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, COLING/ACL 2006*, pp. 841–848, July 2006.

[19] S. Schulz, M. Honeck, and U. Hahn, "Biomedical text retrieval in languages with a complex morphology," in *Proceedings of the meeting of the association for computational linguistics*, vol. 3, pp. 61–68, July 2002.

[20] W. R. Hersh, M. Katherine Crabtree, D. H. Hickam et al., "Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions," *Journal of the American Medical Informatics Association* , vol. 9, no. 3, pp. 283–293, 2002.

[21] J. W. Ely, J. A. Osheroff, M. H. Ebell et al., "Analysis of questions asked by family doctors regarding patient care," *BMJ*, vol. 319, no. 7206, pp. 358–361, 1999.

[22] J. W. Ely, "Obstacles to answering doctors' questions about patient care with evidence: qualitative study," *BMJ*, vol. 324, no. 7339, pp. 710-710.

[23] G. R. Bergus, C. S. Randall, S. D. Sinift, and D. M. Rosenthal, "Does the structure of clinical questions affect the outcome of curbside consultations with specialty colleagues?" *Archives of Family Medicine*, vol. 9, no. 6, pp. 541–547, 2000.

[24] P. Sondhi, P. Raj, V. V. Kumar, and A. Mittal, "Question processing and clustering in INDOC: A biomedical question answering system," *Eurasip Journal on Bioinformatics and Systems Biology*, vol. 2007, 2007.

[25] E. M. Voorhees, "The TREC question answering track," *Natural Language Engineering*, vol. 7, no. 4, pp. 361–378, 2001.

[26] T. Hao, W. Xie, C. Chen, and Y. Shen, "Systematic comparison of question target classification taxonomies towards question answering," *Communications in Computer and Information Science*, vol. 568, pp. 131–143, 2015.

[27] S. J. Athenikos and H. Han, "Biomedical question answering: A survey," *Computer Methods and Programs in Biomedicine*, vol. 99, no. 1, pp. 1–24, 2010.

[28] D. Mollá and J. L. Vicedo, "Question answering in restricted domains: An overview," *Computational Linguistics*, vol. 33, no. 1, pp. 41–61, 2007.

[29] A. Mishra and S. K. Jain, "A survey on question answering systems with classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 28, no. 3, pp. 345–361, 2016.

[30] H. Yu and C. Sable, "Being Erlang Shen: Identifying Answerable Questions," in *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence on Knowledge and Reasoning for Answering Questions*, 2005.

[31] F. Rinaldi, J. Dowdall, G. Schneider, and A. Persidis, "Answering questions in the genomics domain," in *Proceedings of the ACL 2004 Workshop on Question Answering in Restricted Domains*, pp. 46–53, 2004.

[32] A. M. Cohen, J. Yang, S. Fisher, B. Roark, and W. R. Hersh, "The OHSU Biomedical Question Answering System Framework," in *Proceedings of the Sixteenth Text Retrieval Conference*, 2007.

[33] A. B. Abacha and P. Zweigenbaum, "Medical question answering: Translating medical questions into SPARQL queries," in *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, IHI'12*, pp. 41–50, January 2012.

[34] A. Marginean, "Question answering over biomedical linked data with Grammatical Framework," *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 8, no. 4, pp. 565–580, 2017.

[35] A. Ben Abacha and P. Zweigenbaum, "MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies," *Information Processing & Management*, vol. 51, no. 5, pp. 570–594, 2015.

[36] R. M. Terol, P. Martínez-Barco, and M. Palomar, "A knowledge based method for the medical question answering problem," *Computers in Biology and Medicine*, vol. 37, no. 10, pp. 1511–1521, 2007.

[37] H. Zhang, T. W. S. Chow, and Q. M. J. Wu, "Organizing Books and Authors by Multilayer SOM," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 12, pp. 2537–2550, 2016.

[38] H. Zhang, S. Wang, X. Xu, T. W. Chow, and Q. M. Wu, "Tree2Vector: Learning a Vectorial Representation for Tree-Structured Data," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2018.

[39] H. Zhang, S. Wang, Z. Mingbo, X. Xu, and Y. Ye, "Locality Reconstruction Models for Book Representation," *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[40] K.-Y. Chung, J. Yoo, and K. J. Kim, "Recent trends on mobile computing and future networks," *Personal and Ubiquitous Computing*, vol. 18, no. 3, pp. 489–491, 2014.

[41] S.-K. Kang, K.-Y. Chung, and J.-H. Lee, "Development of head detection and tracking systems for visual surveillance," *Personal and Ubiquitous Computing*, vol. 18, no. 3, pp. 515–522, 2014.

[42] H. Jung, K.-Y. Chung, and Y.-H. Lee, "Decision supporting method for chronic disease patients based on mining frequent pattern tree," *Multimedia Tools and Applications*, vol. 74, no. 20, pp. 8979–8991, 2015.

[43] D. Han, M. Lee, and S. Park, "THE-MUSS: Mobile u-health service system," *Computer Methods and Programs in Biomedicine*, vol. 97, no. 2, pp. 178–188, 2010.

[44] S. Wild, G. Roglic, A. Green, R. Sicree, and H. King, "Global prevalence of diabetes: estimates for the year 2000 and projections for 2030," *Diabetes Care*, vol. 27, no. 5, pp. 1047–1053, 2004.

[45] M. E. Cox and D. Edelman, "Tests for screening and diagnosis of type 2 diabetes," *Clinical Diabetes*, vol. 27, no. 4, pp. 132–138, 2009.

[46] G. Danaei, M. M. Finucane, Y. Lu et al., "National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 2·7 million participants," *The Lancet*, vol. 378, no. 9785, pp. 31–40, 2011.

[47] R. A. Bailey, Y. Wang, V. Zhu, and M. F. Rupnow, "Chronic kidney disease in US adults with type 2 diabetes: an updated national estimate of prevalence based on kidney disease: improving Global Outcomes (KDIGO) staging," *BMC Research Notes*, vol. 7, article 415, 2014.

[48] P.-J. Lin, D. M. Kent, A. Winn, J. T. Cohen, and P. J. Neumann, "Multiple chronic conditions in type 2 diabetes mellitus: prevalence and consequences," *The American Journal of Managed Care*, vol. 21, no. 1, pp. e23–e34, 2015.

[49] M. L. Tracey, M. Gilmartin, K. O'Neill et al., "Epidemiology of diabetes and complications among adults in the Republic of Ireland 1998–2015: a systematic review and meta-analysis," *BMC Public Health*, vol. 16, article 132, no. 1, 2016.

[50] L. Yang, J. Shao, Y. Bian et al., "Prevalence of type 2 diabetes mellitus among inland residents in China (2000–2014): A meta-analysis," *Journal of Diabetes Investigation*, vol. 7, no. 6, pp. 845–852, 2016.

[51] K. Iglay, H. Hannachi, P. J. Howie et al., "Prevalence and co-prevalence of comorbidities among patients with type 2 diabetes mellitus," *Current Medical Research and Opinion*, vol. 32, no. 7, pp. 1243–1252, 2016.

[52] P. Zimmet, K. G. Alberti, D. J. Magliano, and P. H. Bennett, "Diabetes mellitus statistics on prevalence and mortality: Facts and fallacies," *Nature Reviews Endocrinology*, vol. 12, no. 10, pp. 616–622, 2016.

[53] I. Dedov, M. Shestakova, M. M. Benedetti, D. Simon, I. Pakhomov, and G. Galstyan, "Prevalence of type 2 diabetes mellitus (T2DM) in the adult Russian population (NATION study)," *Diabetes Research and Clinical Practice*, vol. 115, pp. 90–95, 2016.

[54] K. Waki, H. Fujita, Y. Uchimura et al., "DialBetics: A novel smartphone-based self-management support system for type 2 diabetes patients," *Journal of Diabetes Science and Technology*, vol. 8, no. 2, pp. 209–215, 2014.

[55] T. Hao, H. Liu, and C. Weng, "Valx: A system for extracting and structuring numeric lab test comparison statements from text," *Methods of Information in Medicine*, vol. 55, no. 3, pp. 266–275, 2016.

[56] T. Hao, W. Xie, and F. Xu, "A wordnet expansion-based approach for question targets identification and classification," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 9427, pp. 333–344, 2015.

[57] T. Hao, W. Xie, Q. Wu, H. Weng, and Y. Qu, "Leveraging question target word features through semantic relation expansion for answer type classification," *Knowledge-Based Systems*, vol. 133, pp. 43–52, 2017.

[58] T. Hao, D. Hu, L. Wenyin, and Q. Zeng, "Semantic patterns for user-interactive question answering," *Concurrency and Computation: Practice and Experience*, vol. 20, no. 7, pp. 783–799, 2008.

[59] Z. M. Juan, "An effective similarity measurement for FAQ question answering system," in *Proceedings of the International Conference on Electrical and Control Engineering, ICECE 2010*, pp. 4638–4641, June 2010.

[60] T. Hao and Y. Qu, "QSem: A novel question representation framework for question matching over accumulated question-answer data," *Journal of Information Science*, vol. 42, no. 5, pp. 583–596, 2016.

[61] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the Association for Information Science and Technology*, vol. 41, no. 6, pp. 391–407, 1990.

[62] H. L. Wang and H. Y. Xi, "Chinese synonyms toolkit," 2018, https://github.com/huyingxi/Synonyms.

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 2329876, 10 pages
https://doi.org/10.1155/2018/2329876

Research Article

The Current Status and a New Approach for Chinese Doctors to Obtain Medical Knowledge Using Social Media: A Study of WeChat

Li Liu,[1] Kunyan Wei,[2] Xingting Zhang,[3] Dong Wen,[3] Li Gao,[4] and Jianbo Lei[5,6]

[1]*Sichuan Cancer Hospital & Institute, Chengdu, China*
[2]*Department of Gastroenterology, Affiliated Hospital of Southwest Medical University, Luzhou, China*
[3]*Peking University Third Hospital, Beijing, China*
[4]*Peking University School of Stomatology, Beijing, China*
[5]*Center for Medical Informatics, Peking University, Beijing, China*
[6]*School of Medical Informatics and Engineering, Southwest Medical University, Luzhou, China*

Correspondence should be addressed to Jianbo Lei; jblei@hsc.pku.edu.cn

Received 23 December 2017; Revised 22 March 2018; Accepted 3 April 2018; Published 8 May 2018

Academic Editor: Haoran Xie

Background. WeChat is the most widely and frequently used mobile social media in China and has profoundly integrated into the daily life of many Chinese people. A variety of medicine-related information may be found on WeChat. As users of WeChat, doctors often access health-related information and even provide a variety of medical services or participate in various types of mobile communication with patients. *Objective*. This study is the first attempt to quantitatively explore the approaches by doctors of acquiring medical knowledge using Internet resources especially social media such as WeChat to access knowledge. *Methods*. A self-administered questionnaire was designed, distributed, collected, and analyzed utilizing the online survey tool Sojump. WeChat was adopted to randomly release the questionnaires using snowball sampling and collect the results after a certain amount of time. *Results*. 292 valid questionnaires out of 314 questionnaires by clinical doctors were analyzed. Regarding the current status of accessing medical knowledge among doctors, more than 60% of the doctors regularly used the Internet to search for medical knowledge, 19.86% used WeChat as a channel to acquire medical knowledge, and only 23.97% were satisfied with acquiring medical knowledge through the Internet. Regarding the frequency of WeChat usage, nearly 40% of the doctors accessed WeChat more than 20 times per day and over 70% used WeChat for over half an hour every day. Regarding the status of accessing medical knowledge through WeChat, nearly half (47.26%) of the doctors stated that they often read professional medical articles on WeChat and the most common channel is friends' moment sharing and public account subscriptions, with selection rates of 59.93% and 60.27%, respectively. The most desirable mode of acquiring medical knowledge through WeChat was the following: "professional medical knowledge from peers, with a reminder." *Conclusion*. WeChat has become a nonnegligible means of acquiring medical knowledge for busy Chinese physicians in a mobile environment. Further evaluation and improvement of the quality of medical knowledge on WeChat are needed. The recommendation of individualized articles through social media may become another contributing factor for doctors to acquire medical knowledge effectively and efficiently.

1. Background

The amount of information required for medical practice is growing at an exponential rate, and it has become less practical to completely master the tremendous amount of knowledge with doctors' individual ability [1]. Doctors are faced with the major challenge of handling a flood of professional information. The traditional learning model cannot meet the requirements of the new situation, and there is an urgent need for a new learning model to solve this dilemma [2]. The development of mobile device based social media has provided an idea for a solution [3, 4]. Social media are tools for online learning and distance education. Narrowly speaking, a social medium is a form of discussion by users

based on the website and application, and its main contents are generated by users, who can create, retrieve, and discover these contents [2]. In recent years, social media as learning tools have been growing rapidly in the medical field, forming an online learning mode which is learner-oriented with a relatively high degree of collaboration and participation [5].

The rapid development of the Internet and social media is profoundly changing people's lifestyles. Nearly 90% of American adults are using the Internet [6], of whom 72% access social networking websites [7]. More than 75,000 healthcare professionals around the world share information and discuss treatment options on Twitter [8]. In China, the number of mobile phone users had reached 668 million, and mobile Internet usage has reached 88.9% [9]. WeChat, a free application that provides an instant messaging service to intelligent terminals such as mobile phones launched on January 21, 2011 by Tencent [10]. Via cell phones, WeChat can integrate the functions of instantaneousness and socialization and break through temporal and geographical constraints; thus, it has become the mobile social platform with the largest number of users, the widest application, and the highest frequency in China, showing the most power of influence [11]. According to the latest official data published by Tencent, in September 2015, the number of active WeChat users had reached 570 million, covering over 200 countries in more than 20 languages [12]. WeChat has become not only a cell phone application with many innovative features but also an indispensable tool for everyday use in people's lives, covering more than 90% of China's smart mobile phones and forming a lifestyle in cell phone users [13].

As a social medium, WeChat carries a variety of topics and articles, of which many are concerning medical expertise [14–16]. Doctors are a part of the large community of WeChat users, and many researchers have noted this phenomenon. Medical education by social media has been accepted by doctors around the world [5, 17–19]. In recent years, an increasing number of studies on emerging social media in the medical field, from mobile phone texts to Twitter, Facebook, and microblogs, have been reported [20–22]. However, few quantitative studies on doctors' attitude in using social media are available in the literature [23]. A study on the top website for doctors in the US reported that many doctors seek advice from the social networking website when confronted with a difficult case and discuss health policy and medical research online [24]. However, studies on the use of social media in physicians for continuing medical education are also lacking. How social media with strong interaction and high usage such as WeChat affect the behavior, learning of medical knowledge, doctor-patient interaction, and even online medical services in physicians is an area that warrants investigation. This study targets continuing medical education and is the first to quantitatively investigate Chinese doctors' attitudes toward the channels for acquiring medical knowledge and continuing medical education using WeChat.

2. Research Methods

2.1. Questionnaire Design. Regarding the survey, Sojump .com (URL: https://www.sojump.com/) was used to generate, distribute, and collect the questionnaire. Sojump.com is a professional platform for online surveys, assessment, and voting; it focuses on providing a series of powerful and user-friendly services, including online questionnaire design, data collection, customized reports, and analysis of the survey results. Compared to the traditional method of investigation and other survey sites or survey systems, Sojump.com has the obvious advantages of being rapid, easy to use, and inexpensive [25].

Since there is no existing questionnaire able to answer the current status of physicians to acquire medical knowledge via social media, especially WeChat, we have to develop our own questionnaire. According to the objective of this study, a multidisciplinary team was formed to review relevant literature and discuss study needs and methods to develop the questionnaire. A final questionnaire on the knowledge, attitude, and behavior regarding WeChat and health and medical problems was designed based on the experience of previous questionnaire development and a consideration of the characteristics of social media. The survey mainly included the following framework: general demographic indicators, the usage situation of WeChat, the status of obtaining medical knowledge and health education through WeChat, the desired approach for health education, and the existing problems of healthcare and medical information on WeChat. A pilot test was conducted for the initial questionnaire after being developed by a team of three doctors and two survey experts. Based on the type of related questions, the length of the questions, the answer options, and the time to complete the questionnaire in the feedback, the questionnaire was modified and improved, resulting in the final questionnaire, which was available online at the following URL: www.sojump.com/jq/5854804.aspx.

2.2. Data Sample and Survey Methods. WeChat is a social tool with a powerful user group in China, and nearly half of its active users have more than 100 WeChat friends [12]. In this study, the questionnaires were distributed by WeChat with the snowball sampling method. From October 10, 2015 to October 21, 2015, the self-designed questionnaire for medical doctors on using WeChat to acquire medical knowledge was distributed to 2000 friends and a variety of groups belonging to the author's WeChat account (approximately 20 groups, with a minimum of eight members and a maximum of 314 members). Each receiver was asked to forward the questionnaire link to other WeChat users through friends' network on WeChat, and ultimately, 314 questionnaires were collected; of these, 292 questionnaires from the physician respondents were valid.

2.3. Data Processing. The collected questionnaires were analyzed, and the invalid questionnaires were excluded based on the following exclusion criteria: repeated IP address in the response; logical errors in the answers; the same answer for successive questions; completion time being too short; and completion time being too long. A total of 22 invalid questionnaires were excluded, resulting in 292 valid questionnaires. Data analysis was conducted using the built-in statistical function in the backend of Sojump.com.

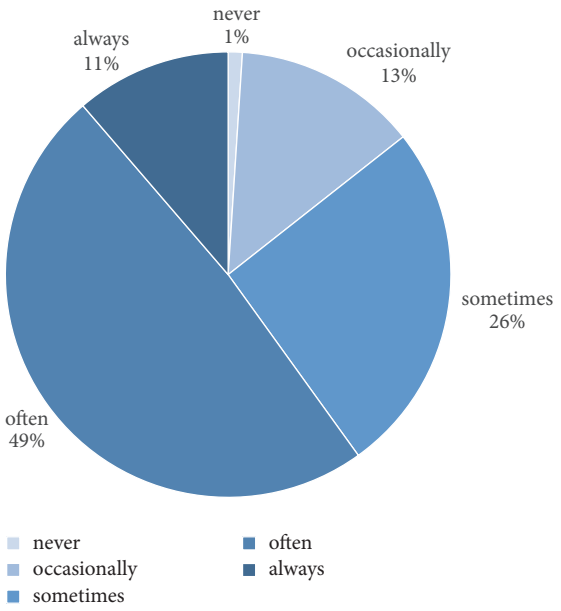


Figure 1: Frequency of Internet usage by doctors to search for medical knowledge.

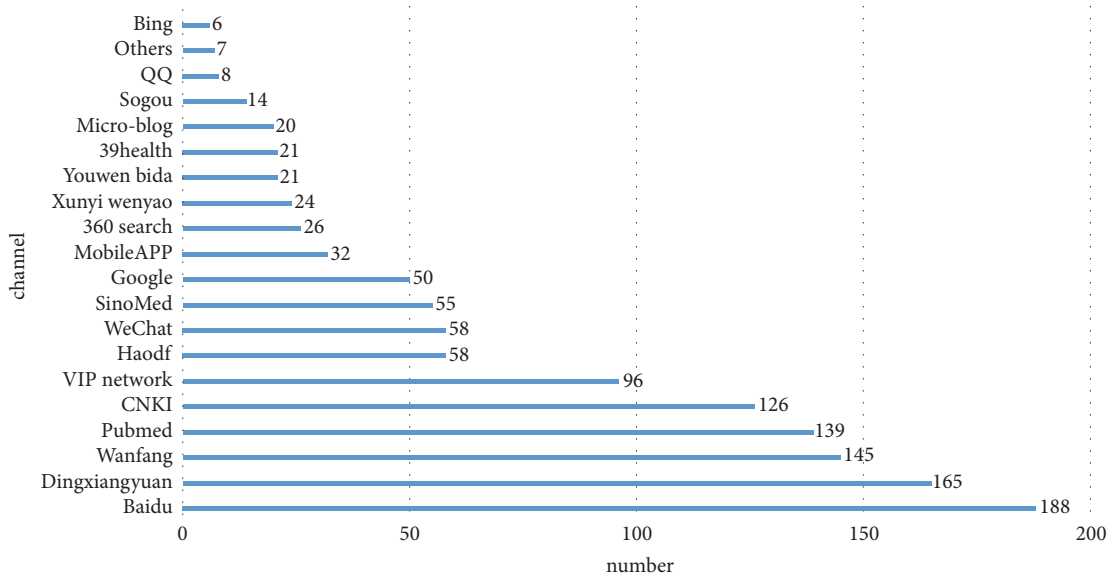


Figure 2: The approaches by doctors to acquire medical knowledge by searching the Internet.

3. Results

3.1. General Information. A total of 292 valid questionnaires were collected from the respondents, including 134 males and 158 females, with ages ranging from 18 to 50. Of the respondents, all were clinicians with 173 who had the educational level of a master's degree or above, accounting for 59.24% of all respondents. They covered all major clinical departments of the hospital. Their titles were evenly distributed from residents to chief physician. According to the monitoring results of the respondents' IP addresses, the geographic distribution of the respondents covered most areas of China, with the most respondents being in Beijing. All questionnaires were submitted by mobile phone. Table 1 shows the demographic details of the respondents.

3.2. Investigation of the Status of Acquiring Professional Medical Knowledge via the Internet in the Doctors. In this study, more than half of the doctors (59.93%) regularly used the Internet to search for medical knowledge; 142 respondents (48.63%) often used it, and 33 respondents (11.30%) always used it, as shown in Figure 1. The most popular search engine was Baidu (with a selection rate of 64.38%, 188/292), and 19.86% physicians selected WeChat as an approach to acquire specialized knowledge, as shown in Figure 2. Doctors' satisfaction with using the Internet for acquiring medical

Table 1: Summary of general information.

Category	Number of respondents	Percentage (%)
Gender		
Males	134	45.89
Females	158	54.11
Age		
Under 18	0	0
18–25	48	16.44
26–30	61	20.89
31–40	95	32.53
41–50	72	24.66
51–60	15	5.14
Above 60	1	0.34
Education		
High school/vocational school	2	0.68
College/university	117	40.07
Master	119	40.75
Doctorate or above	54	18.49
Professional level		
Intern	48	16.44
Resident	69	23.63
Attending doctor	84	28.77
Associate Professor	61	20.89
Professor	30	10.27
Total	292	

knowledge was not high. Only 23.97% of the respondents were satisfied in the overall evaluation of searching the Internet for medical expertise, as shown in Figure 3.

3.3. Survey of WeChat Usage. Regarding the usage of WeChat, 285 respondents accessed WeChat every day, accounting for 97.60% of the sample. These daily users were classified by the number of times accessing WeChat, with an interval of 10. The number of respondents who accessed it 1–10 times per day was the highest, accounting for 32.28% (92/285), followed by respondents who accessed it 10–20 times, accounting for 29.47% (84/285). Notably, the number of doctors who accessed it more than 20 times per day accounted for 38.24% (45 + 27 + 37 = 109, 109/285), indicating that the frequency of WeChat usage in Chinese doctors was very high, as shown in Figure 4. In addition, the doctors also spent a long period of time using WeChat. Figure 5 shows that over 70% (206/285) of the respondents spent more than 30 minutes per day using WeChat.

3.4. Survey of the Situation for Acquiring Professional Medical Knowledge Using WeChat. Nearly half (47.26%, 111 + 27 = 138, and 138/292) of the respondents agreed that they often

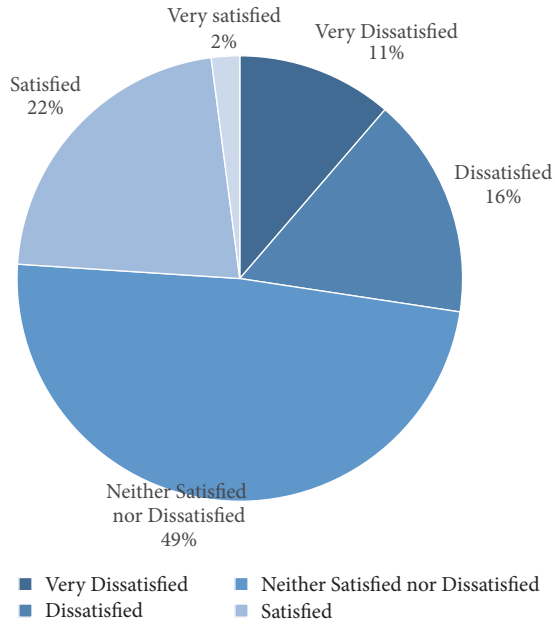


Figure 3: Satisfaction with using the Internet to search for medical knowledge.

found medical articles on WeChat, as shown in Figure 6. Additionally, 41.44% of the respondents said that they regularly accessed professional medical articles on WeChat. When the doctors were searching for medical information on WeChat, the most common methods included browsing friends' moments (with a selection rate of 59.93%, 175/292) and public account subscriptions (with a selection rate of 60.27%, 176/292), followed by Dingxiangyuan (an online community for Chinese physicians), WeChat search and group chat, as shown in Figure 7. Readability was the highest in the doctors' assessment of medical knowledge on WeChat, whereas the ratings for professionalism and usefulness were relatively low, as shown in Figure 8. Currently, the potential problems of WeChat with regard to medical expertise of the highest concern for the doctors included the homogenization of information, unguaranteed professionalism, and too many advertisements, with selection rates of 67.81%, 66.44%, and 61.30%, respectively. The issue over which there was the least concern was readability. 45.21% of respondents believed that medical knowledge from WeChat had no help; 36.30% of respondents were neutral; only 18.50% believed that WeChat health information could improve professional skills.

3.5. The Most Desirable Means to Acquire Medical Knowledge. Study participants held high expectations for new media like WeChat. In the survey of the most desirable channel by which to acquire medical knowledge through the Internet, Baidu was still the preferred website for searching (with a selection rate of 55.82%, 163/292), and the selection rate for WeChat was 34.25% (100/292), as shown in Figure 9. Regarding the doctors' mode of acquiring medical knowledge through WeChat, "professional medical knowledge from peers, with a reminder," had the highest score of 80.67 out of 100 points, indicating that professional knowledge posted

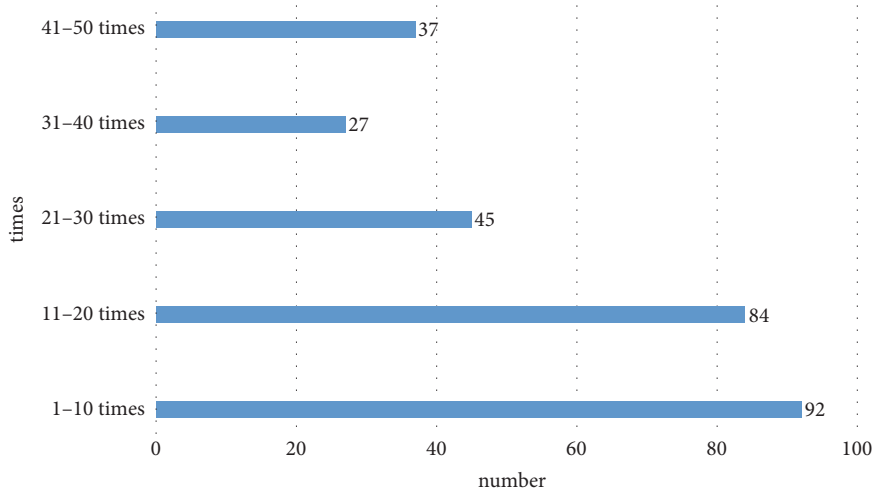


Figure 4: The number of times accessing WeChat per day.

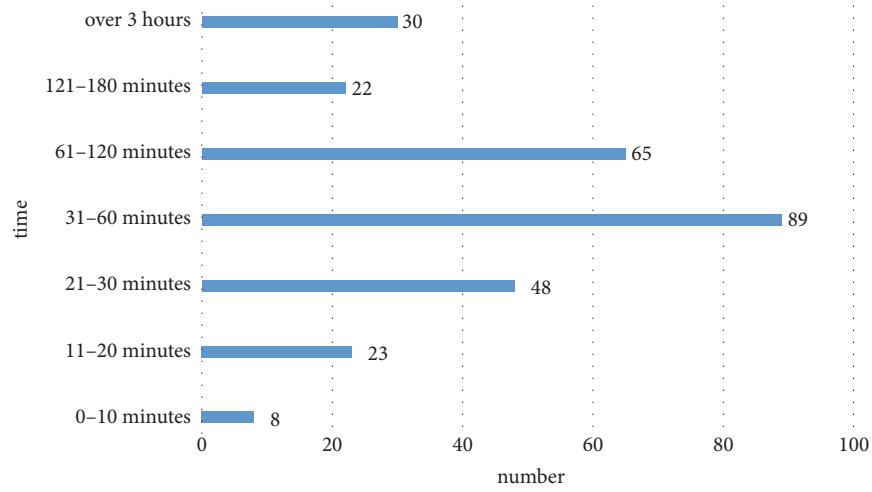


Figure 5: Length of time of WeChat usage with the number of users.

through interaction between peers was recognized the most by the doctors. Details were shown in Figure 10.

4. Discussion

4.1. WeChat Has Become an Important and Nonnegligible Mobile Supplement for Doctors to Obtain Medical Knowledge. Continuing medical education plays an irreplaceable role in the education of medical professionals. Well-developed continuing medical education and training for healthcare professionals after the completion of their academic education in college can allow them to continuously update their professional knowledge and improve their operational capacity throughout their career, meet the needs of the development of medical science and technology, and enhance the overall level of medical care. Due to the particularity of the contents in continuing medical education, although medical professionals are eager to learn with clear objectives, they have different knowledge backgrounds, and their disposable time is significantly random and fragmented, lacking a long fixed period of time for concentrated study. The effectiveness of the implementation of mobile continuing medical education has not been very satisfactory.

WeChat is a potential way of acquiring professional medical knowledge for busy doctors. As is known to all, many people nowadays use social media as alternative source of information. Because the Internet makes retrieval easy and there are no time and space constraints, it is recognized by the vast majority of doctors and has increasingly become the mode of acquiring medical knowledge with the highest degree of trust in the doctors in the sample. However, only 23.97% of them were satisfied with acquiring medical knowledge through the Internet. It is noteworthy that, in this study, 19.86% of the physicians selected the new interactive social media WeChat as the Internet search application for acquiring medical knowledge. Over 70% of the surveyed physicians spent more than 30 minutes per day on WeChat, occupying an important portion of the doctors' fragmented

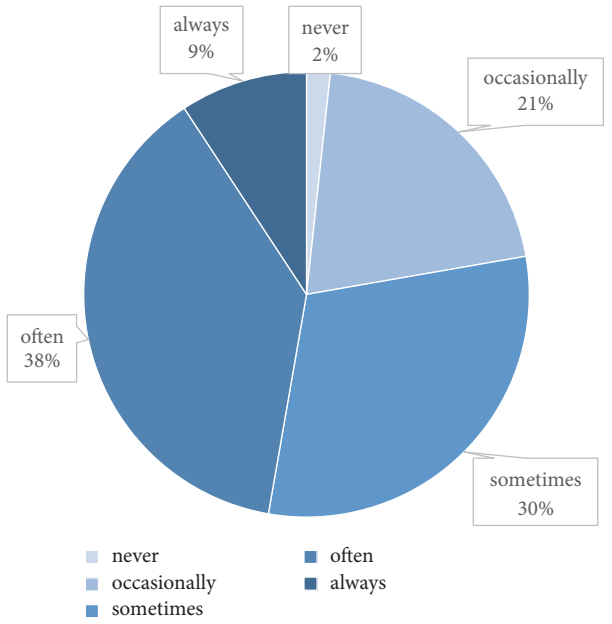


Figure 6: Frequency of accessing medical knowledge on WeChat.

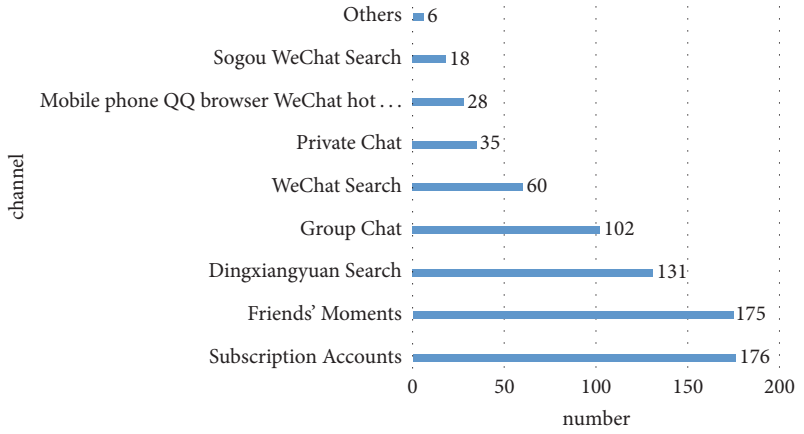


Figure 7: Approaches to searching for medical knowledge on WeChat.

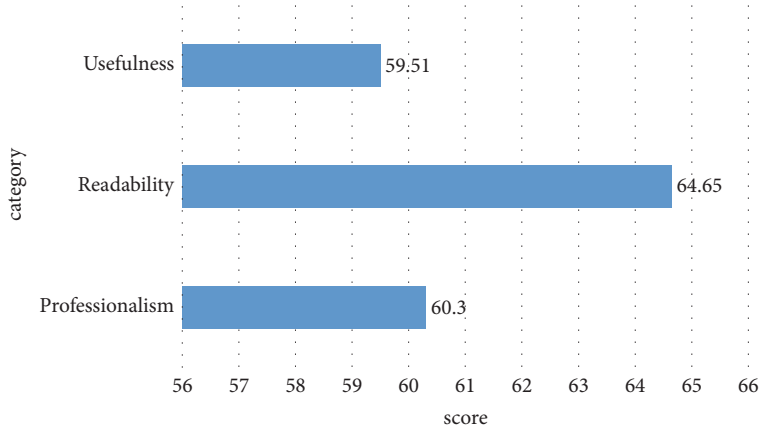


Figure 8: Quality assessment of medical knowledge on WeChat by doctors.

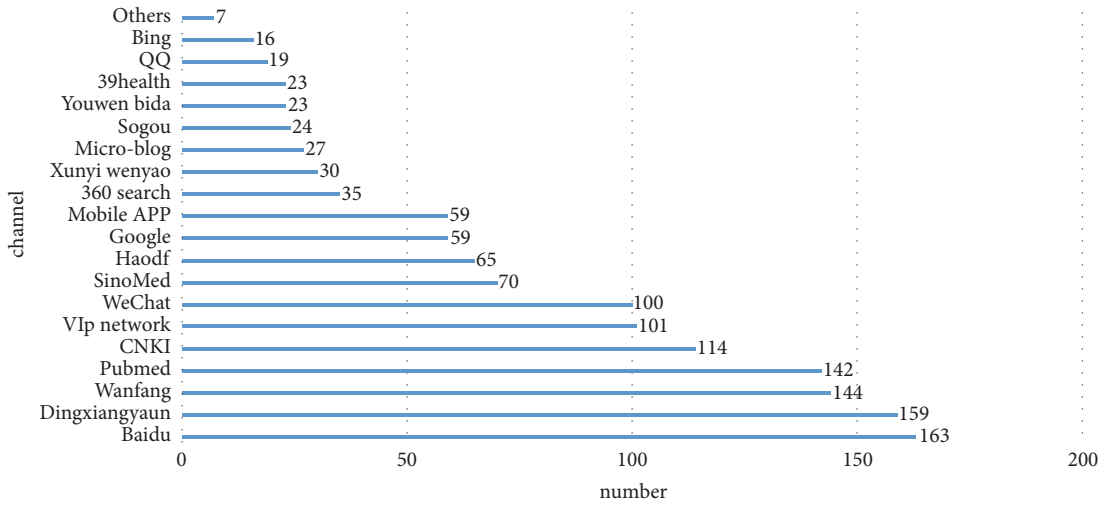


Figure 9: The most desirable channels by which to acquire medical knowledge.

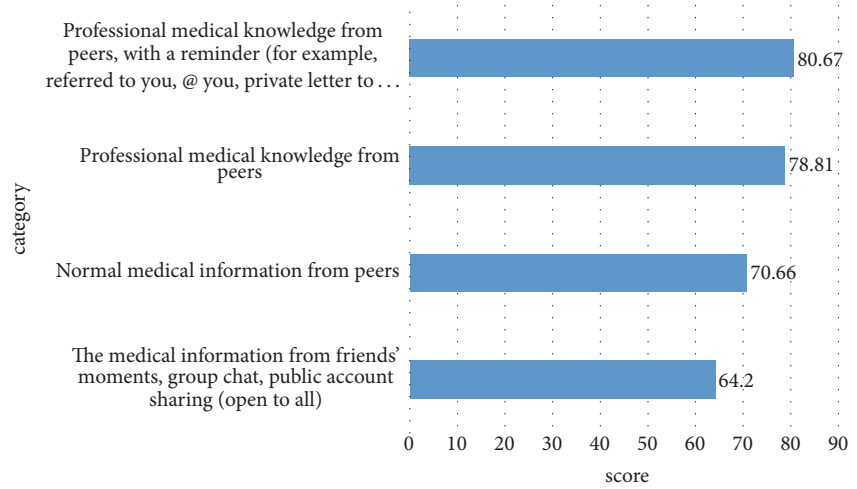


Figure 10: Mode for acquiring medical knowledge on WeChat.

time. It has become an important supplement for the doctors to obtain medical knowledge.

Studies have corroborated the feasibility of WeChat as a tool for medical education. The research and practice of Bai et al., who used WeChat to construct a management platform, also confirmed that the use of WeChat not only meets the daily habits of residents but also expands teaching resources to a certain extent, thereby achieving the exchange and sharing of excellent teaching resources [26]. Wang et al. analyzed the application of mobile learning tools in e-Learning and indicated that the application of WeChat and Line in mobile learning can enhance the communication of information and contribute to information sharing and problem solving [27]. The study by Li et al. shows that WeChat can improve the quality and capacity of doctors, greatly expand their purview of clinical diseases, and promote the exchange of knowledge of different cases among resident doctors from different departments and disciplines [28].

WeChat and other social media education platforms have been widely recognized by the contents of medical education due to their characteristics related to flexible knowledge dissemination, freedom from the constraints of time and place, the large amount of information, rapid dissemination, the large audience size, ease of management, and strong social attributes and interaction, and they have become an important platform for medical education organizations to perform medical education. Currently, acquiring medical knowledge and the latest medical information through the Internet and new media has become a main means through which active medical staff access information [29, 30].

4.2. Improving Physicians' Awareness of the Various Searching Methods of WeChat Can Help Them Better Use WeChat to Access Professional Knowledge. Doctors may not be familiar with the various searching methods on WeChat. We found that the approaches of acquiring medical knowledge using WeChat by doctors focused on browsing friends' moments, public account subscriptions, Dingxiangyuan browsing for WeChat, and group chat (with selection rates of 59.93%, 60.27%, 44.86%, and 34.93%, resp.). The selection rates for

WeChat search, Sogou (a third-party search engine) search for WeChat, the mobile phone QQ browser, and WeChat hot articles were relatively low, but these methods are actually able to greatly enhance the efficiency of WeChat search. Thus, it is necessary to strengthen training on the relevant skills, thus providing greater convenience for doctors to take full advantage of WeChat in acquiring sufficient medical knowledge. Information overload and the proliferation of false information are the important issues of new media. How to obtain valuable information from redundant information and how to discern truth from falsehood through independent judgment have become the new challenge for all individuals in the era of new media. Whether the user has the abilities to select, understand, question, assess, create, produce, and speculatively respond to the targeted information, particularly the ability of independent judgment, has become a prerequisite that determines whether the user can benefit from the age of new media [31]. With Chinese doctors, due to the unitary structure of professional knowledge and the relative lack of information technology, it is more difficult for them to obtain valuable information from new media, which seriously affects the effectiveness of new media education. Therefore, it is particularly important to enhance the training level in new media knowledge and to improve the new media literacy of doctors [32].

4.3. Assessing and Improving the Professionalism of Medical Knowledge on WeChat Is the Future Direction for Application and Research. There were major concerns with the quality of medical information on WeChat, especially with regard to content. Doctors acquire medical knowledge on the premise of its quality. Only correct professional knowledge has value in learning. In the assessment of the quality of medical knowledge on WeChat by the doctor respondents, the rating on readability was the highest, whereas the evaluation ratings on professionalism and professional improvement were lower. Currently, the potential problems of medical information on WeChat of the highest concern for the doctors included the homogenization of information, unguaranteed professionalism, and too many advertisements. The issue of the least concern was readability, and 45.21% respondents considered that it had no significant impact on improving their professional skills. This result indicates that nearly half of the doctors distrusted the medical knowledge on WeChat. The overabundant flood of homogeneous information, advertisements, and unprofessional information without filtering may be the main reasons. In a study from another country, Decamp reached a similar conclusion [33]. Because the current medical knowledge on WeChat is mostly from public accounts, whose operators are mostly not professional authorities, the quality of professional medical knowledge on WeChat cannot be guaranteed, which has become the most worrisome problem in using WeChat as a tool for continuing education for doctors. Therefore, it is essential to evaluate article quality and improve the dissemination of high-quality articles. Due to the high professionalism of medical knowledge and its significant impact on public health [34, 35], how to more accurately assess the quality of medical knowledge on WeChat, how to ensure the quality of medical knowledge, and how to promote the dissemination of high-quality articles on WeChat warrant further study because these factors will determine the frequency and effectiveness of the future application of WeChat in the continuing medical education of doctors.

4.4. Limitations of This Study. First, the present study used the random snowball sampling method in the survey, and the only respondents were doctors in the private WeChat friends' circle of the researchers. Although the doctors were distributed nationwide and the responses were random and independent, these samples are insufficient to represent all doctors in China. However, they may reflect the situation of acquiring medical knowledge and WeChat usage in the doctors of China to a certain extent. This study is the first large-scale study in China. Second, because the questionnaires were distributed online, they were convenient in many aspects including questionnaire publishing, collection, and analysis. However, the questionnaire collection period was short and had a small number of valid responses (only 292); thus, randomness may exist in the responses. These factors most likely caused the bias in the results; thus, this study is only an exploratory attempt in the field. Accurate and more convincing research requires a further expanded sample size, a strictly random method, improvement in the reliability and validity of the questionnaire, and strict control of all the details in the survey so that the results can be more objective and scientific.

4.5. Directions and Trends for Future Study. This study is only a preliminary investigation from the perspective of doctors' acquisition of medical knowledge. Our further research is being conducted as follows: first, a larger and strictly randomized investigation on the channel of continuing education in doctors and the impact of WeChat on continuing education; second, a study of other behaviors and the effects of WeChat usage by doctors, such as research on the doctor-patient interaction and doctors' attitudes toward and approach to online medical consultation service using WeChat; and third, research on quality assessment methods for medical knowledge on WeChat and how to achieve more effective communication for high-quality articles through WeChat. In summary, social media, especially WeChat, have profoundly changed the public's lifestyle and become part of the public's everyday life. WeChat will certainly have an increasingly important impact on healthcare services, such as medical knowledge learning, health education, and medical services in doctors. Research and application in this area will be a very important direction.

5. Conclusion

The traditional methods of acquiring medical knowledge have been unable to meet the needs of doctors. The present study shows that WeChat has already become an important mobile means for doctors to acquire medical expertise. Naturally, doctors have not yet fully mastered the approaches to performing searches on WeChat, and the quality of medical knowledge on WeChat is worrisome. Thus, further quality assessment methods are needed to assess and improve the

quality of medical knowledge on WeChat. These factors preclude doctors from taking advantage of WeChat for professional development. In conclusion, WeChat plays an increasingly important role in acquiring medical knowledge and continuing education for Chinese doctors.

Disclosure

In addition, this paper was selected as one of the outstanding papers and partially presented in conference CHIP in November 2017.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Jianbo Lei developed the conceptual framework and research protocol for the study. Xingting Zhang and Li Gao conducted the questionnaire creation, distribution, collection, and data analysis. Li Liu and Kunyan Wei interpreted the data and drafted the manuscript. Jianbo Lei made major revisions. All authors approve the final version of the manuscript. Li Liu and Kunyan Wei contributed equally to this work.

Acknowledgments

This study was partly supported by the National Natural Science Foundation of China (NSFC) (Grant nos. 81471756 and 81771937).

References

[1] A. P. Abernethy, L. M. Etheredge, P. A. Ganz et al., "Rapid-learning system for cancer care," *Journal of Clinical Oncology*, vol. 28, no. 27, pp. 4268–4274, 2010.

[2] B. S. McGowan, M. Wasko, B. S. Vartabedian, R. S. Miller, D. D. Freiherr, and M. Abdolrasulnia, "Understanding the factors that influence the adoption and meaningful use of social media by physicians to share medical information," *Journal of Medical Internet Research*, vol. 14, no. 5, article no. e117, 2012.

[3] L. Dini, C. Galanski, S. Döpfmer et al., "Online platform as a tool to support postgraduate training in general practice – A case report," *GMS Journal for Medical Education*, vol. 34, no. 5, Article ID Doc59, 2017.

[4] S. X. Li and R. Pinto-Powell, "Revisiting the merits of a mandatory large group classroom learning format: an MD-MBA perspective," *Medical Education Online*, vol. 22, no. 1, p. 1396174, 2017.

[5] S. Batt-Rawden, T. Flickinger, J. Weiner, C. Cheston, and M. Chisolm, "The role of social media in clinical excellence," *The Clinical Teacher*, vol. 11, no. 4, pp. 264–269, 2014.

[6] L. Rainie, "A biography of the pew research center's Internet & American life project," *Encyclopedia of Cyber Behavior*, vol. 1, pp. 25–41, 2012.

[7] Brenner J, S., Internet & American Life Project Tracking Survey. Pew Research Center, 2013.

[8] J. M. Alpert and F. E. Womble, "Just What the Doctor Tweeted: Physicians' Challenges and Rewards of Using Twitter," *Health Communication*, vol. 31, no. 7, pp. 824–832, 2016.

[9] China Internet Network Information Center, the 37th statistic reports on China's Internet development, 2015. http://www.cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/201601/t20160122_53271.htm.

[10] Tencent.Inc.Wechat home page. http://www.wechat.com/en/.

[11] Baidu Encyclolpedia, Introduction to WeChat, http://baike.baidu.com/link?url=LLNl376yqNHpkwQKcZfROHtw_bMvb-g150T7o4-ou6lMoT2PV3LeT761vREB2JEnYXhUzjhVIod7P4-p32O1Cl6DrX9DmpTkCPEQV_d4gc5My.

[12] Curiositychina, statistic report of WeChat users, http://curiositychina.com/.

[13] H. Tong, *The broadcast of and impact of WeChat*, vol. 09, Chongqing Social Sciences, 2013.

[14] C. He, S. Wu, Y. Zhao et al., "Social Media–Promoted Weight Loss Among an Occupational Population: Cohort Study Using a WeChat Mobile Phone App-Based Campaign," *Journal of Medical Internet Research*, vol. 19, no. 10, p. e357, 2017.

[15] X. Zhang, D. Wen, J. Liang, and J. Lei, "How the public uses social media wechat to obtain health information in China: A survey study," *BMC Medical Informatics and Decision Making*, vol. 17, article no. 66, 2017.

[16] B. Cao, C. Liu, M. Durvasula et al., "Social media engagement and HIV testing among men who have sex with men in China: A nationwide cross-sectional survey," *Journal of Medical Internet Research*, vol. 19, no. 7, article no. e251, 2017.

[17] T. S. O'Hagan, D. Roy, B. Anton, and M. S. Chisolm, "Social Media Use in Psychiatric Graduate Medical Education: Where We Are and the Places We Could Go," *Academic Psychiatry*, vol. 40, no. 1, pp. 131–135, 2016.

[18] L. N. Ko, J. Rana, and S. Burgin, "Teaching & Learning Tips 5: Making lectures more "active"," *International Journal of Dermatology*, vol. 57, no. 3, pp. 351–354, 2018.

[19] D. Rozgonjuk, K. Saal, and K. Täht, "Problematic Smartphone Use, Deep and Surface Approaches to Learning, and Social Media Use in Lectures," *International Journal of Environmental Research and Public Health*, vol. 15, no. 1, p. 92, 2018.

[20] L. Nicolai, M. Schmidbauer, M. Gradell et al., "Facebook groups as a powerful and dynamic tool in medical education: Mixed-method study," *Journal of Medical Internet Research*, vol. 19, no. 12, article no. e408, 2017.

[21] L. Sinnenberg, A. M. Buttenheim, K. Padrez, C. Mancheno, L. Ungar, and R. M. Merchant, "Twitter as a tool for health research: A systematic review," *American Journal of Public Health*, vol. 107, no. 1, pp. e1–e8, 2017.

[22] R. C. Chang, H. Lu, P. Yang, and P. Luarn, "Reciprocal Reinforcement Between Wearable Activity Trackers and Social Network Services in Influencing Physical Activity Behaviors," *JMIR mHealth and uHealth*, vol. 4, no. 3, p. e84, 2016.

[23] F. Gholami-Kordkheili, V. Wild, and D. Strech, "The impact of social media on medical professionalism: A systematic qualitative review of challenges and opportunities," *Journal of Medical Internet Research*, vol. 15, no. 8, article no. e184, 2013.

[24] N. S. Fogelson, Z. A. Rubin, and K. A. Ault, "Beyond likes and tweets: An in-depth look at the physician social media landscape," *Clinical Obstetrics and Gynecology*, vol. 56, no. 3, pp. 495–508, 2013.

[25] About sojump. https://www.wjx.cn/html/aboutus.aspx.

[26] Y. Bai, M. Xu, S. Chen, and X. Zhao, *Feasibility study of building a cardiovascular teaching platform using WeChat*, vol. 3, Northwest Medical Education, 2015.

[27] J. Wang, W. C. W. Yu, and E. Wu, "Empowering Mobile Assisted Social E-Learning: Students' Expectations and Perceptions," *World Journal of Education*, vol. 3, 2013.

[28] L. Li, K. Lv, and S. Zhu, *Application of WeChat information release system based on mobile terminal in standardized training of burn surgery residents*, vol. 5, Northwest Medical Education, 2014.

[29] E. M. Geyer and D. E. Irish, "Isolated to integrated: An evolving medical informatics curriculum," *Medical Reference Services Quarterly*, vol. 27, no. 4, pp. 451–461, 2008.

[30] E. S. Schwenk, L. F. Chu, R. K. Gupta, and E. R. Mariano, "How Social Media Is Changing the Practice of Regional Anesthesiology," *Current Anesthesiology Reports*, vol. 7, no. 2, pp. 238–245, 2017.

[31] B. M. Walter, R. M. Schmid, and S. Von Delius, "Improving patient information - Are the new media already requested? - A questionnaire study at a gastroenterology outpatient clinic," *Zeitschrift für Gastroenterologie*, vol. 55, no. 6, pp. 551–556, 2017.

[32] J. Salem, H. Borgmann, A. MacNeily et al., "New Media for Educating Urology Residents: An Interview Study in Canada and Germany," *Journal of Surgical Education*, vol. 74, no. 3, pp. 495–502, 2017.

[33] M. Decamp, "Physicians, social media, and conflict of interest," *Journal of General Internal Medicine*, vol. 28, no. 2, pp. 299–303, 2013.

[34] Y. Sugawara, H. Narimatsu, A. Hozawa, L. Shao, K. Otani, and A. Fukao, "Cancer patients on Twitter: A novel patient community on social media," *BMC Research Notes*, p. 699, 2012.

[35] K. C. Chretien, J. Azar, and T. Kind, "Physicians on twitter," *Journal of the American Medical Association*, vol. 305, no. 6, pp. 566–568, 2011.

Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 2379208, 8 pages
https://doi.org/10.1155/2018/2379208

Research Article

Recognizing Continuous and Discontinuous Adverse Drug Reaction Mentions from Social Media Using LSTM-CRF

Buzhou Tang, Jianglu Hu, Xiaolong Wang, and Qingcai Chen

Key Laboratory of Network Oriented Intelligent Computation, Shenzhen Graduate School of Harbin Institute of Technology, Shenzhen 518055, China

Correspondence should be addressed to Buzhou Tang; tangbuzhou@gmail.com

Received 26 December 2017; Accepted 13 March 2018; Published 19 April 2018

Academic Editor: Tianyong Hao

Social media in medicine, where patients can express their personal treatment experiences by personal computers and mobile devices, usually contains plenty of useful medical information, such as adverse drug reactions (ADRs); mining this useful medical information from social media has attracted more and more attention from researchers. In this study, we propose a deep neural network (called LSTM-CRF) combining long short-term memory (LSTM) neural networks (a type of recurrent neural networks) and conditional random fields (CRFs) to recognize ADR mentions from social media in medicine and investigate the effects of three factors on ADR mention recognition. The three factors are as follows: (1) representation for continuous and discontinuous ADR mentions: two novel representations, that is, "BIOHD" and "Multilabel," are compared; (2) subject of posts: each post has a subject (i.e., drug here); and (3) external knowledge bases. Experiments conducted on a benchmark corpus, that is, CADEC, show that LSTM-CRF achieves better *F*-score than CRF; "Multilabel" is better in representing continuous and discontinuous ADR mentions than "BIOHD"; both subjects of comments and external knowledge bases are individually beneficial to ADR mention recognition. To the best of our knowledge, this is the first time to investigate deep neural networks to mine continuous and discontinuous ADRs from social media.

1. Introduction

With rapid growth of online health social networks, such as DailyStrength.com [1], Askapatient.com [2], MedHelp.org [3], and PatientsLikeMe.com [4], more and more patients share their personal health-related information through social media posts. This information can be utilized for public health monitoring, particularly for pharmacovigilance via mining adverse drug reactions (ADRs) using natural language processing techniques.

ADRs, noxious and unintended responses to medicinal products occurring at doses normally used in man for the prophylaxis, diagnosis or therapy of disease or for the restoration, and correction or modification of physiological function [5] are an essential part of drug safety. There usually are two kinds of mechanisms to discover ADRs: (1) some of ADRs are discovered during phase III clinical trials for drug development and (2) some are revealed during postmarketing surveillance. In the case of the postmarketing surveillance, traditionally, ADRs are identified by individual patients and their physicians using official adverse event reporting systems (AERS). In recent years, there have been some attempts at mining ADRs and drug-drug interactions [6, 7] from unstructured text in clinical records, literature, and health-related social media, and their experimental results have shown the effectiveness of mining ADRs from unstructured text. In this study, we focus on recognizing ADR mentions, including continuous and discontinuous ADR mentions, from social media according to all comments to specific drugs. For example, given a user's post "I still have pain in arms and legs with much stiffness." our goal is to extract three ADR mentions, namely, "pain in arms," "pain in. . .legs," and "stiffness," where "pain in arms" and "stiffness" are continuous ADR mentions composed of continuous words and "pain in. . .legs" is a discontinuous ADR mention composed of discontinuous words.

Although there have been a number of methods proposed for recognizing ADR mentions from social media, most of

them used traditional machine learning methods such as conditional random fields (CRFs) to recognize continuous ADR mentions. In this work, we propose a deep neural network (called LSTM-CRF), which combines long short-term memory (LSTM) neural networks (a type of recurrent neural networks, RNNs) and conditional random fields (CRFs), to recognize both continuous and discontinuous ADR mentions from social media. Compared to CRFs, the advantages of LSTM-CRF lie in that LSTM neural networks have strong expressive ability to capture long context without time-intensive feature engineering. It has shown better performance than CRFs for some sequence labeling tasks such as part-of-speech (POS) tagging and named entity recognition (NER) [8].

In order to comprehensively investigate LSTM-CRF on ADR mention recognition, we compare two novel unified representations (i.e., "BIOHD" and "Multilabel") for continuous and discontinuous ADR mentions and studied effects of post subjects and external knowledge bases. All models are evaluated on a benchmark corpus, that is, CADEC, composed of 1250 forum posts for 12 drugs taken from AskaPatient.com, where each post has been manually annotated with ADR mentions. Our results show that LSTM-CRF performs better than CRF, "Multilabel" is more suitable than "BIOHD" to represent continuous and discontinuous ADR mentions, and both subject of posts and external knowledge bases are individually beneficial to ADR mention recognition.

On the whole, the contributions of this work can be summarized as follows: (1) we introduce a deep neural network that combines LSTM neural networks and CRFs to recognize continuous and discontinuous ADR mentions at first time; (2) we compare two unified representations for continuous and discontinuous ADR mentions; (3) we investigate the effects of post subjects and knowledge bases; (4) we conduct empirical evaluation of all models on a benchmark corpus.

This paper is organized as follows: in Section 2, we survey related work; Section 3 introduces the LSTM-CRF model; and Section 4 depicts our experiments in detail; we provide discussion in Section 5 and Section 6 concludes the paper. An earlier version of the paper has been presented in The 3rd China Health Information Processing Conference (CHIP-2017).

2. Related Work

In recent years, social media has been increasingly used for medical research, especially for pharmacovigilance via mining ADRs from health-related posts. Certain quantities of studies have been proposed for ADR mention recognition from corpus construction to methods. Posts from DailyStrength.com [1], Yahoo Wellness Groups [9], Askapatient.com [2], Medications.com [10], WebMD.com [11], MedHelp.org [3], SteadyHelath.com [12], PatientsLikeMe.com [4], parenting websites [13], various disease-specific forums such as Diabetes and Cancer [14], Twitter [15], Facebook [16], and other websites or forums [17] have been collected to mine ADRs. On these data, varieties of methods have been employed to recognize ADR mentions. They may fall into three categories: lexicon-based [18–29], pattern-based [30, 31], and machine learning-based [32, 33]. The earliest work, the pioneering work of Leaman et al. in 2010 [18], utilized a lexicon-based method to recognize ADR mentions from user posts regarding six drugs from DailyStrength.com. In this work, 450 out of 3600 posts were used for system development and the remanding post for system evaluation. Although lexicon-based methods can successfully recognize ADR mentions using several extensive and available ADR resources, they cannot address challenges such as idiomatic expressions and misspelled mentions. To conquer some of them, pattern-based methods over lexicon-based methods were proposed to detect inexact-match ADR mentions. For example, Yates et al. [34] designed seven patterns to recognize ADR mentions from posts regarding five drugs for breast cancer from Askpatient.com, Drugs.com, and DrugRatingZ.com. The limitation of pattern-based methods is the need for large amounts of data to generate patterns. Recently, with some annotation data available, machine learning-based methods, such as CRFs, are becoming more and more popular with promising performances, where ADR mention recognition is considered as a sequence-labeling problem. Sarker and Gonzalez [35] made a comprehensive review of text mining techniques for ADR mining before 2015. As mentioned in this review, most state-of-the-art machine learning methods are based on CRFs with rich hand-crafted features, and only continuous ADR mentions are taken into account. In 2016, Pacific Symposium on Biocomputing (PSB) launched a shared task on mining social media to exploit natural language processing techniques for ADR extraction in tweets, where subtask 2 is to automatically extract ADR mentions in user posts. In this subtask, only continuous ADR mentions were considered, and machine learning methods based on CRFs achieved best results again [33].

Actually, discontinuous entity mentions are very common in the medical domain. As reported in [36], discontinuous disorder mentions in clinical text accounted for about 10%. In social media, discontinuous ADR mentions also usually appear. Karimi et al. [37] annotated a corpus of adverse drug events including both continuous and discontinuous ADR mentions, that is, CADEC. To recognize continuous and discontinuous ADR mentions simultaneously, Metke-Jimenez and Karimi [38] followed Tang et al.'s [36] way to represent them in a unified schema and used CRFs with baseline features, including bag-of-words, character n-grams, and word shapes. To the best of our knowledge, this is the only study that considered both continuous and discontinuous ADR mentions in the task of ADR mention recognition. There are also some other studies conducted on this corpus; however, all of them only consider continuous ADR mentions or convert every discontinuous ADR mention into one or more continuous ADR mentions. For example, Tutubalina and Nikolenko [39] proposed a method, uniting RNNs and CRFs, to recognize ADR mentions on CADEC. They excluded overlaps between spans of discontinuous ADRs by selecting the longest continuous span and combining these ADRs into a single continuous ADR.

Table 1: Examples of continuous and discontinuous ADR mentions represented by "BIOHD" and "Multilabel," respectively.

Sentence 1	I	experienced	severe	pain	in	my	left	shoulder	.			
BIOHD	O	O	O	DB	DI	O	DI	DI	O			
Multilabel	O	O	O	B	I	O	I	I	O			
Sentence 2	I	still	have	pain	in	arms	and	legs	with	much	stiffness	.
BIOHD	O	O	O	HB	HI	DB	O	DB	O	O	B	O
Multilabel	O	O	O	B	I	I	O	O	O	O	B	O
				B	I	O		I			O	

Deep learning methods have been increasingly applied to solve NLP tasks in the medical domain and achieve better performance than CRFs. In the case of ADR mention recognition, Stanovsky et al. [40] employed RNNs with word embeddings trained on a Blekko medical corpus in conjunction with entity embeddings trained on DBpedia. If an entity mention was a lexical match with one of DBpedia entities, then the entity embeddings trained on DBpedia replaced word embeddings of all words in the entity mention. Tutubalina and Nikolenko [39] utilized multilayer RNNs (LSTM and GRU) with CRFs and achieved better performance than RNNs and CRFs individually. However, no study focuses on applying deep learning methods to recognize both continuous and discontinuous ADR mentions simultaneously.

3. Methods

Before recognizing continuous and discontinuous ADR mentions, we should know how to represent them. Therefore, in this section, we introduce representation schemas for both continuous and discontinuous ADR mentions at first and then machine learning methods.

3.1. Representations. Two novel representations are adopted in our study: "BIOHD" and "Multilabel." "BIOHD" is an extension of traditionally named entity representation schema "BIO" (B-beginning of a ADR mention, I-inside of a ADR mention, O-outside of a ADR mention) by introducing two additional tags: "H," a part shared by multiple medical mentions (e.g., ADR mentions), and "D," a part of a discontinuous medical mention not shared by other mentions. "Multilabel" allows a token to be labeled with more than one tag, and each tag corresponds to the position in one mention. The number of tags a token has is determined by how many mentions it appears in. Table 1 gives us examples of continuous and discontinuous ADR mentions represented by "BIOHD" and "Multilabel," respectively. In sentence 1, there is one discontinuous ADR mention "pain in. . .left shoulder," while there are two continuous ADR mentions, "stiffness" and "pain in arms," and one discontinuous ADR mention, "pain in. . .legs," in sentence 2. The ADR mentions "pain in arms" and "pain in. . .legs" in sentence 2 share the part "pain in," For convenience, a token's multiple tags can be combined into one tag. For example, sentence 2 can be tagged with "I/O still/O have/O pain/B-B in/I-I arms/I-O and/O legs/O-I with/O much/O stiffness/B-O ./O," where each token and its tag(s) are separated by "/," and multiple tags are joined by "-." In this study, the maximum number of tags a token has is set to 4 according to the statistic results from the training corpus.

3.2. LSTM-CRF. When continuous and discontinuous ADR mentions are represented by "BIOHD" or "Multilabel," recognizing them still can be formulated as a sequence labeling problem. In this study, we use LSTM-CRF to model this problem. Figure 1 illustrates the architecture of LSTM-CRF, which is composed of three layers as follows: (1) input layer, (2) LSTM-layer, and (3) CRF layer.

The input layer takes in different types of embeddings of each token. The embeddings used in this study include word embeddings, char-level embeddings, subject-related embeddings, and knowledge-based embeddings.

The LSTM layer uses bidirectional LSTM neural networks to generate hidden context representation at each position. Given a sentence $s = w_1 w_2 \cdots w_n$, where each word w_t $(1 \le t \le n)$ is represented by x_t (i.e., concatenation of word embeddings, char-level embeddings, subject-related embeddings, and knowledge-based embeddings of word w_t), the bidirectional LSTM neural networks take a sequence of embeddings $x = x_1 x_2 \cdots x_n$ as input and output a sequence of hidden context representations $h = h_1 h_2 \cdots h_n$, where $h_t = [h_{ft}^{\mathrm{T}}, h_{bt}^{\mathrm{T}}]^{\mathrm{T}}$ $(1 \le t \le n)$ is a concatenation of the outputs of both forward and backward LSTM neural networks.

The CRF layer takes a sequence of hidden context representations $h = h_1 h_2 \cdots h_n$ as input, estimates the probabilities of label sequences (from a predefined set), and returns the one of highest probability. As shown in [41], the conditional probability of a label sequence $y = y_1 y_2 \cdots y_n$ for $h = h_1 h_2 \cdots h_n$ is computed as follows (only taking the first-order linear chain CRF as an example here):

$$p_{\lambda,\mu}(y \mid h) = \frac{\exp\left(\sum_{t=1}^{n}\left(\lambda_{y_{t-1}y_t} + \left\langle \mu_{y_t}, h_t \right\rangle\right)\right)}{\exp\left(\sum_{y' \in Y(h)} \sum_{t=1}^{n}\left(\lambda_{y'_{t-1}y'_t} + \left\langle \mu_{y'_t}, h_t \right\rangle\right)\right)}, \quad (1)$$

where $\lambda_{y_{t-1}y_t}$ is the transform score for $y_{t-1}y_t$, $\langle \mu_{y_t}, h_t \rangle$ is the emission score generating y_t given h_t, and $Y(h)$ represents all possible label sequences for h.

4. Experiments

4.1. Corpus. We use a publicly available annotated corpus called CSIRO Adverse Drug Event Corpus (CADEC) from

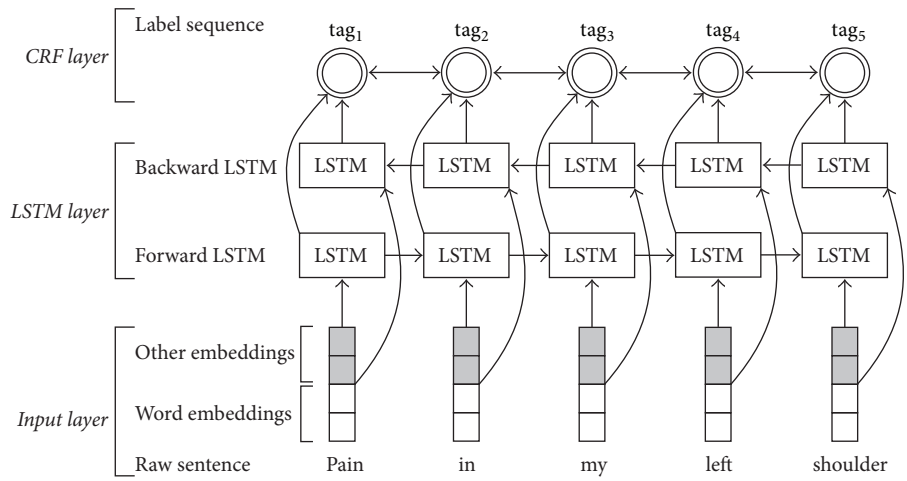


Figure 1: Architecture of LSTM-CRF for continuous and discontinuous ADR mentions recognition. "tag_i" is the label for the ith token defined in "BIOHD" or "Multilabel".

AskaPatient.com to evaluate the performance of LSTM-CRF. On AskaPatient.com, all comments are grouped by drugs. CADEC contains 1250 posts about 12 drugs {Voltaren, Cataflam, Voltaren-XR, Arthrotec, Pennsaid, Solaraze, Flector, Cambia, Zipsor, Diclofenac Sodium, Diclofenac Potassium, and Lipitor}, and the posts are manually annotated with five types of ADR-related events {ADR, Drug, Disease, Symptom, and Findings}. In CADEC, there are 6318 ADR mentions, 1000 out of which are discontinuous ADR mentions, accounting for 15.83%. Among the 1000 discontinuous ADR mentions, 918 share some parts with others, that is, overlapping.

4.2. Evaluation Metrics. We use precision (P), recall (R), F-score (F), and accuracy (Acc) to evaluate ADR mention recognition system. They are defined as follows:

$$\begin{aligned}
\text{precision} &= \frac{n_{TP}}{n_{TP} + n_{FP}}, \\
\text{recall} &= \frac{n_{TP}}{n_{TP} + n_{FN}}, \\
F\text{-score} &= \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \\
\text{accuracy} &= \frac{n_{TP} + n_{TN}}{n_{TP} + n_{FN} + n_{FP} + n_{TN}},
\end{aligned} \tag{2}$$

where n_{TP} is the number of ADR mentions correctly predicted by a system, n_{FP} is the number of ADR mentions predicted by a system but not in the gold standard corpus, n_{FN} is the number of ADR mentions in the gold standard corpus but not predicted by a system, and n_{TN} is the number of nonentities (tagged as "O") correctly predicted.

Two kinds of criteria, that is, strict and relaxed, are adopted to calculate P, R, F, and Acc. The strict criterion refers to the fact that an ADR mention is correctly predicted only when it is exactly the same as the gold-standard one. The relaxed criterion refers to the fact that an ADR mention is correctly predicted as long as it overlaps with the gold-standard one.

4.3. Experimental Results. We reimplement Metke-Jimenez and Karimi's CRF-based system [38], select LSTM-CRF only using word embeddings and char-level embeddings as a baseline, compare LSTM-CRF with CRF, and investigate effects of ADR mention representations, post subjects, and knowledge bases on LSTM-CRF. The word embeddings are initialized by GloVe [42] and 100-dimensional pretrained embeddings on a large-scale unlabeled dataset from Wikipedia [43]. We use bidirectional LSTM neural networks to extract char-level embeddings. The LSTM neural networks take a character sequence of each word (each char is represented by character embeddings) as input and output two hidden sequence representations. The last two outputs of the bidirectional LSTM neural networks are simply concatenated into char-level embeddings. The character embeddings are randomly initialized from uniform distribution ranging in $[-1, 1]$, and its dimension is set to 25. The dimension of the char-level embeddings is also set to 25. The subject-related embeddings are randomly initialized from uniform distribution ranging in $[-1, 1]$, and their dimension is set to 10. We label each token in a sentence with BIOES (B-beginning of a ADR mention, I-inside of a ADR mention, O-outside of a ADR mention, E-end of a ADR mention, and S-a single-token ADR mention) through knowledge-based looking up, and utilize 10-dimensional embeddings, randomly initialized from uniform distribution ranging in $[-1, 1]$, to represent each token's label. The SIDER database (http://sideeffects.embl.de/) and ADR lexicon (http://diego.asu.edu/downloads/publications/ADRMine/ADR_lexicon.tsv) are two knowledge bases used in this study. All embeddings are fine-tuned during training.

As there is no fixed way to divide CADEC into two parts, one for system development and the other one for system evaluation, we adopt 10-fold cross-validation. Following the previous study [8], we set other parameters of LSTM-CRF as follows: dimension of LSTM hidden states: 100, optimizer: SGD, learning rate: 0.005, dropout rate: 0.5, and maximum

number of epochs: 200. The results of different methods are shown in Table 2, where the highest values are highlighted in bold.

Table 2 shows that the methods using "Multilabel" outperform that using "BIOHD." For example, the strict F-score of CRF using "Multilabel" is higher than CRF using "BIOHD" by 0.76% (0.6060 versus 0.5984). Compared with CRF, LSTM-CRF only using word embeddings and char-level embeddings (denoted as LSTM-CRF (baseline)) achieves better performance. When continuous and discontinuous ADR mentions are represented by "BIOHD," LSTM-CRF achieves higher strict F-score than CRF by 4.92% (0.6476 versus 0.5984), while when continuous and discontinuous ADR mentions are represented by "Multilabel," LSTM-CRF achieves higher strict F-score than CRF by 4.99% (0.6559 versus 0.6060). Both subject-based embeddings (denoted by "subject" in Table 2) and knowledge-based embeddings (denoted by "knowledge" in Table 2) are individually beneficial to LSTM-CRF. When subject-based embeddings are added, the strict F-score of LSTM-CRF using "Multilabel" is improved from 0.6559 to 0.6636, which is the highest F-score. When knowledge-based embeddings are added, the strict F-score of LSTM-CRF using "Multilabel" is improved from 0.6559 to 0.6593. When both of them are added, LSTM-CRF achieves a strict F-score of 0.6614, which is a slightly lower than the highest one (i.e., 0.6636). The differences between strict F-scores and relaxed F-scores of the same methods exceed 20%, indicating that exactly detecting ADR mentions' boundaries is not easy.

In addition, we also analyze the performances of different methods on continuous and discontinuous ADR mentions, respectively, as shown in Table 3, where the highest indices are highlighted in bold. LSTM-CRF using "Multilabel" representation and subject-based embeddings achieves the highest strict F-score of 69.94% for continuous ADR mentions, while LSTM-CRF using "Multilabel" representation and knowledge-based embeddings achieves the highest strict F-score of 41.87% for discontinuous ADR mentions. The difference between the two highest F-scores is near 20%. The baseline LSTM-CRF achieves much higher F-scores than CRF on continuous ADR mentions by about 5% and on discontinuous ADR mentions by about 8%. The methods using "Multilabel" almost always outperform that using "BIOHD" on continuous ADR mention recognition. The strict F-score difference between the methods using the two different representations ranges from 0.2% to 0.59%. For discontinuous ADR mentions, the methods using "Multilabel" always outperform that using "BIOHD" by 4.93% in average strict F-score, which is much larger than the strict F-score difference between the methods using the two different representations for continuous ADR mentions.

5. Discussion

In this study, we propose a deep neural network (i.e., LSTM-CRF) to recognize continuous and discontinuous ADR mentions from medical social media, compare it with CRF, and investigate the effects of different factors on the proposed method. Similar to other related tasks such as NER and POS tagging, LSTM-CRF outperforms CRF on continuous and discontinuous ADR mention recognition. The methods using "Multilabel" outperform that using "BIOHD." Both subject-based embeddings and knowledge-based embeddings are individually beneficial to ADR mention recognition, but when both of them are simultaneously added, the performance is not further improved.

The reason why the methods using "Multilabel" outperform that using "BIOHD" may lie in that "Multilabel" has better representation ability than "BIOHD," especially for discontinuous ADR mentions. In theory, "Multilabel" is perfect (with a coverage of 100%), while "BIOHD" is imperfect [36]. The coverage of "BIOHD" on CADEC is 89.36%. Because of this, the strict F-score difference between systems using "Multilabel" and "BIOHD" for discontinuous ADR mentions is much larger than that for continuous ADR mention, although the distributions of continuous and discontinuous ADR mentions also affect the performance. For example, there are four ADR mentions, "Extremely bad pains in hands," "Extremely bad pains in…arms," "Extremely bad pains in…muscles," and "Extremely bad pains in…quivering," recognized by "BIOHD" in "Extremely/HB bad/HI pains/HI in/HI hands/DB ,/O arms/DB ,/O and/O muscles/DB are/O constantly/O quivering/DB ./O"; however, in fact, there are only three ADR mentions, "Extremely bad pains in hands," "Extremely bad pains in…arms," and "muscles…quivering." Since different drugs have different ADRs, adding subject-based embeddings amounts to adding the relations between drugs and their ADRs, similar to relations in knowledge bases. It may be the reason for the improvement from subject-based embeddings and why simultaneously adding both the subject-based embeddings and knowledge-based embeddings does not bring further improvements.

As the distributions of continuous and discontinuous ADR mentions are imbalanced, it is easy to understand that the strict F-score difference of the same methods for continuous and discontinuous ADR mentions is not small. How to tackle data imbalance is a possible direction for further improvement, which will be considered in the future.

Although LSTM-CRF shows much better performance than CRF, the performance of LSTM-CRF is not very good, indicating that recognizing continuous and discontinuous ADR mentions from medical social media is still challenging. The main challenge is exactly determining all words or tokens of mentions, not some of them. The errors of LSTM may fall into the following three categories. (1) Some modifiers are missing. For example, there are three continuous ADR mentions, "long time flatulence," "Achilles tendon tightness," and "dizziness," in post "long time flatulence, Achilles tendon tightness, and dizziness." The first one is wrongly recognized as "flatulence." (2) Some discontinuous ADR mentions are wrongly recognized as continuous mentions by combining words or tokens between all parts. For example, the discontinuous ADR mention "hair…thinning" in sentence "I took this drug a few years ago and went off it because my hair started thinning." is wrongly recognized as a continuous ADR mention "hair started thinning." (3) There are some combination errors between continuous ADR mentions and

Table 2: Results of LSTM-CRF and CRF on CADEC.

Method	Representation	Strict				Relaxed			
		P	*R*	*F*	Acc	*P*	*R*	*F*	Acc
CRF	BIOHD	0.6707 ± 0.035	0.5402 ± 0.026	0.5984 ± 0.029	0.7037 ± 0.017	0.9175 ± 0.025	0.7700 ± 0.019	0.8373 ± 0.019	0.8892 ± 0.011
	Multilabel	0.6865 ± 0.038	0.5424 ± 0.028	0.6060 ± 0.032	0.7101 ± 0.019	0.9177 ± 0.025	0.7702 ± 0.018	0.8375 ± 0.017	0.8848 ± 0.009
LSTM-CRF (baseline)	BIOHD	0.6496 ± 0.045	0.6456 ± 0.034	0.6476 ± 0.033	0.7398 ± 0.02	0.8796 ± 0.025	**0.9213** ± 0.02	0.9000 ± 0.017	0.9346 ± 0.014
	Multilabel	0.6616 ± 0.037	0.6502 ± 0.035	0.6559 ± 0.034	0.7467 ± 0.02	0.8880 ± 0.032	0.9106 ± 0.014	0.8991 ± 0.019	0.9301 ± 0.021
LSTM-CRF + subject	BIOHD	0.6571 ± 0.037	0.6540 ± 0.029	0.6556 ± 0.032	0.7451 ± 0.019	**0.8983** ± 0.024	0.9141 ± 0.019	**0.9061** ± 0.016	**0.9501** ± 0.025
	Multilabel	**0.6780** ± 0.037	0.6499 ± 0.035	**0.6636** ± 0.033	**0.7522** ± 0.019	0.8928 ± 0.033	0.9167 ± 0.014	0.9046 ± 0.019	0.9321 ± 0.015
LSTM-CRF + knowledge	BIOHD	0.6438 ± 0.046	**0.6632** ± 0.034	0.6534 ± 0.038	0.7444 ± 0.023	0.8835 ± 0.028	0.9164 ± 0.013	0.8997 ± 0.017	0.9488 ± 0.023
	Multilabel	0.6682 ± 0.039	0.6505 ± 0.031	0.6593 ± 0.033	0.7490 ± 0.018	0.8858 ± 0.04	0.9139 ± 0.01	0.8996 ± 0.019	0.9291 ± 0.021
LSTM-CRF + all	BIOHD	0.6567 ± 0.036	0.6592 ± 0.03	0.6580 ± 0.031	0.7465 ± 0.017	0.8869 ± 0.03	0.9212 ± 0.019	0.9037 ± 0.017	0.9424 ± 0.015
	Multilabel	0.6633 ± 0.041	0.6594 ± 0.027	0.6614 ± 0.032	0.7514 ± 0.019	0.8896 ± 0.03	**0.9213** ± 0.01	0.9052 ± 0.018	0.9337 ± 0.011

Table 3: Performances of different methods on continuous and discontinuous ADR mentions, respectively (using strict criterion).

Method	Representation	Continuous ADR mention			Discontinuous ADR mention		
		P	*R*	*F*	*P*	*R*	*F*
CRF	BIOHD	0.6882	0.5950	0.6382	0.3093	0.2495	0.2762
	Multilabel	0.6896	0.5993	0.6413	0.5217	0.2405	0.3293
LSTM-CRF (baseline)	BIOHD	0.6720	0.7002	0.6858	0.3579	0.3563	0.3571
	Multilabel	0.6703	0.7062	0.6877	0.4825	0.3533	0.4079
LSTM-CRF + subject	BIOHD	0.6800	0.7092	0.6943	0.3573	0.3613	0.3593
	Multilabel	**0.6882**	0.7109	**0.6994**	0.4400	0.3263	0.3747
LSTM-CRF + knowledge	BIOHD	0.6698	0.7165	0.6924	0.3417	**0.3802**	0.3600
	Multilabel	0.6753	0.7062	0.6904	**0.5096**	0.3553	**0.4187**
LSTM-CRF + all	BIOHD	0.6813	**0.7173**	0.6988	0.3406	0.3513	0.3459
	Multilabel	0.6718	0.7154	0.6929	0.4841	0.3623	0.4144

discontinuous ADR mentions. For example, there are three ADR mentions, "Severe pain in buttocks," "Severe pain in. . .left leg," and "sciatica like symptoms," in "Severe pain in buttocks and left leg sciatica like symptoms." The last two mentions are wrongly recognized as "left leg sciatica like symptoms." Some of these errors may be corrected by using structures of sentences. It is another case of our future work. The proposed representations actually provide two ways to connect different parts of discontinues entities; therefore, the proposed methods may have potential use for relation extraction, such as drug-drug interaction extraction.

6. Conclusions

In this paper, we investigate deep neural network-based ADR mention recognition. A deep neural network (called LSTM-CRF) combining long short-term memory (LSTM) neural networks (a type of recurrent neural networks) and conditional random fields (CRFs) is proposed to recognize continuous and discontinuous ADR mentions from social media in medicine and analyze effects of ADR mention representations, subject-based embeddings, and knowledge-based embeddings. Experiments conducted on a benchmark corpus show that (1) LSTM-CRF outperforms CRF; (2) "Multilabel" representation is more suitable for continuous and discontinuous ADR mention recognition than "BIOHD"; (3) both subject-based embeddings and knowledge-based embeddings are individually beneficial for continuous and discontinuous ADR mention recognition. Moreover, some possible directions for further improvement are also presented.

Conflicts of Interest

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

Acknowledgments

This paper is supported in part by the following grants: National 863 Program of China (2015AA015405), National Natural Science Foundation of China (NSFC) (61573118, 61402128, 61473101, and 61472428), Special Foundation for Technology Research Program of Guangdong Province (2015B010131010), Strategic Emerging Industry Development Special Funds of Shenzhen (JCYJ20160531192358466 and JCYJ20170307150528934), Innovation Fund of Harbin Institute of Technology (HIT.NSRIF.2017052), and CCF-Tencent Open Research Fund (RAGR20160102).

References

[1] "Online Support Groups and Forums at DailyStrength," http://www.dailystrength.org.

[2] "Ask a Patients," http://www.askapatient.com.

[3] "MedHelp Medical Support Communities," http://www.medhelp.org/forums/list.

[4] "PatientsLikeMe: live better, together," http://www.patientslikeme.com.

[5] Guideline ICHHT, "Guideline for good clinical practice," *Journal of Postgraduate Medicine*, vol. 47, no. 1, pp. 45–50, 2001.

[6] I. Segura-Bedmar, P. Martínez, and M. H. Zazo, "Semeval-2013 task 9: extraction of drug-drug interactions from biomedical texts (ddiextraction 2013)," in *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Second Joint Conference on Lexical and Computational Semantics (∗SEM)*, vol. 2, pp. 341–350, 2013.

[7] I. Segura-Bedmar, P. Martínez, and D. Sánchez-Cisneros, "The 1st DDIExtraction-2011 challenge task: extraction of drug-drug interactions from biomedical texts," in *Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction*, pp. 1–9, September 2011.

[8] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," *Computation and Language*, 2015.

[9] "Yahoo! Groups," https://groups.yahoo.com/neo.

[10] "The premier community to talk about health," http://www.medications.com/.

[11] WebMD, "Better information. Better health," https://www.webmd.com/.

[12] "SteadyHealth – ask, share, contribute," http://www.steadyhealth.com/.

[13] "Parenting.co.uk - How You Can Be a Better Parent," http://www.parenting.co.uk/.

[14] "Diabetes and Cancer," http://www.diabetes.co.uk/diabetes-complications/diabetes-and-cancer.html.

[15] "Twitter. It's what's happening," https://twitter.com/.

[16] "Facebook," https://www.facebook.com/.

[17] Y. Zeng, X. Liu, Y. Wang et al., "Recommending education materials for diabetic questions using information retrieval approaches," *Journal of Medical Internet Research*, vol. 19, no. 10, 2017.

[18] R. Leaman, L. Wojtulewicz, R. Sullivan et al., "owards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks," in *Proceedings of the 2010 workshop on biomedical natural language processing*, pp. 117–125, Association for Computational Linguistics, 2010.

[19] A. Benton, L. Ungar, S. Hill et al., "Identifying potential adverse effects using the web: A new approach to medical hypothesis generation," *Journal of Biomedical Informatics*, vol. 44, no. 6, pp. 989–996, 2011.

[20] J. Hadzi-Puric and J. Grmusa, "Automatic drug adverse reaction discovery from parenting websites using disproportionality methods," in *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pp. 792–797, IEEE Computer Society, 2012.

[21] C. C. Yang, H. Yang, L. Jiang, and M. Zhang, "Social media mining for drug safety signal detection," in *ProcEedings of the 2012 International Workshop on Smart Health and Wellbeing*, pp. 33–40, 2012.

[22] X. Liu and H. Chen, "AZDrugMiner: an information extraction system for mining patient-reported adverse drug events in online patient forums," in *International Conference on Smart Health*, pp. 134–150, Springer, Berlin, Germany.

[23] S. Yeleswarapu, A. Rao, T. Joseph, V. G. Saipradeep, and R. Srinivasan, "A pipeline to extract drug-adverse event pairs from multiple data sources," *BMC Medical Informatics and Decision Making*, vol. 14, no. 1, article no. 13, 2014.

[24] C. C. Freifeld, J. S. Brownstein, C. M. Menone et al., "Digital drug safety surveillance: Monitoring pharmaceutical products in Twitter," *Drug Safety*, vol. 37, no. 5, pp. 343–350, 2014.

[25] I. Segura-Bedmar, R. Revert, and P. Martínez, "Detecting drugs and adverse events from Spanish health social media streams," in *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@ EACL*, pp. 106–115, 2014.

[26] X. Liu, J. Liu, and H. Chen, "Identifying adverse drug events from health social media: A case study on heart disease discussion forums," in *International Conference on Smart Health*, pp. 25–36, Springer, Cham, Switzerland, 2014.

[27] K. O'Connor, P. Pimpalkhute, A. Nikfarjam, R. Ginn, K. L. Smith, and G. Gonzalez, "Pharmacovigilance on twitter? Mining tweets for adverse drug reactions," in *AMIA Annual Symposium Proceedings*, vol. 2014, artricle 924, American Medical Informatics Association, 2014.

[28] C. C. Yang, H. Yang, and L. Jiang, "Postmarketing drug safety surveillance using publicly available health-consumer-contributed content in social media," *ACM Transactions on Management Information Systems (TMIS)*, vol. 5, no. 1, article no. 2, 2014.

[29] H. Sampathkumar, X.-W. Chen, and B. Luo, "Mining adverse drug reactions from online healthcare forums using Hidden Markov Model," *BMC Medical Informatics and Decision Making*, vol. 14, no. 1, article 91, 2014.

[30] A. Nikfarjam and G. H. Gonzalez, "Pattern mining for extraction of mentions of Adverse Drug Reactions from user comments," in *AMIA Annual Symposium Proceedings*, vol. 2011, article 1019, American Medical Informatics Association, 2011.

[31] A. Yates and N. Goharian, "ADRTrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites," in *European Conference on Information Retrieval*, pp. 816–819, Springer, Berlin, Germany, 2013.

[32] K. Jiang and Y. Zheng, "Mining Twitter data for potential drug effects," in *International Conference on Advanced Data Mining and Applications*, pp. 434–443, Springer, Berlin, Germany, 2013.

[33] A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, and G. Gonzalez, "Pharmacovigilance from social media: Mining adverse drug reaction mentions using sequence labeling with word embedding cluster features," *Journal of the American Medical Informatics Association* , vol. 22, no. 3, pp. 671–681, 2015.

[34] A. Yates, N. Goharian, and O. Frieder, "Extracting adverse drug reactions from social media," in *AAAI Conference on Artificial Intelligence*, pp. 2460–2467, 2015.

[35] A. Sarker and G. Gonzalez, "Portable automatic text classification for adverse drug reaction detection via multi-corpus training," *Journal of Biomedical Informatics*, vol. 53, pp. 196–207, 2015.

[36] B. Tang, Q. Chen, X. Wang et al., "Recognizing disjoint clinical concepts in clinical text using machine learning-based methods," in *AMIA Annual Symposium Proceedings*, vol. 2015, article 1184, American Medical Informatics Association, 2015.

[37] S. Karimi, A. Metke-Jimenez, M. Kemp, and C. Wang, "Cadec: A corpus of adverse drug event annotations," *Journal of Biomedical Informatics*, vol. 55, pp. 73–81, 2015.

[38] A. Metke-Jimenez and S. Karimi, "Concept extraction to identify adverse drug reactions in medical forums: a comparison of algorithms," *Artificial Intelligence*, 2015.

[39] E. Tutubalina and S. Nikolenko, "Combination of deep recurrent neural networks and conditional random fields for extracting adverse drug reactions from user reviews," *Journal of Healthcare Engineering*, vol. 2017, Article ID 9451342, 9 pages, 2017.

[40] G. Stanovsky, D. Gruhl, and P. N. Mendes, "Recognizing mentions of adverse drug reaction in social media using knowledge-infused recurrent models," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, vol. 1, pp. 142–151, 2017.

[41] J. Lafferty, A. McCallum, and F. C. N. Pereira, "CondiTional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

[42] J. Pennington, R. Socher, and C. D. Manning, "GloVe: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, vol. 14, pp. 1532–1543, 2014.

[43] "Wikipedia," https://www.wikipedia.org/.