PROCEEDINGS OF SCIENCE

# Social Network Recommendation Algorithm based on ICIP

**Lin Bi** [1]
*School of Computer Science and Technology, Changchun University of Science and Technology*
*Changchun, 130022, China*
*E-mail:* `bilin7080@163.com`

**Xiaoqiang Di**[2]
*School of Computer Science and Technology, Changchun University of Science and Technology*
*Changchun, 130022, China*
*E-mail:* `dixiaoqiang@126.com`

**Weiwu Ren**[3]
*School of Computer Science and Technology, Changchun University of Science and Technology*
*Changchun, 130022, China*
*E-mail:* `renweiwu@cust.edu.cn`

**Ying Zhang** [4]
*Information Center，Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences*
*Changchun, 130022, China*
*E-mail:* `starsugar_zy@163.com`

The recent years has witnessed rapid development of social network platforms. To enable timely and effective selection of information that is valuable and raises interests, a variety of personalized recommendation algorithms have been put into practice. The improved clustering based on the Isolation Point (ICIP) algorithm is presented based on clustering. In order to overcome the shortcomings of the traditional clustering, the process of Isolation Point has been included in the ICIP algorithm. In this paper, the ICIP algorithm is used for topic extraction of WeChat articles. According to the characteristics of the social network platform, the data noise reduction and modeling is adopted first and then text classification is achieved based on the similarity. The ICIP is applied to remove isolated points, improve clustering accuracy and reduce noise. Compared with other clustering algorithms, the ICIP algorithm has higher accuracy and efficiency.

*ISCC2017*
*16-17 December 2017*
*Guangzhou, China*

[1] Speaker

[2] This work was financially supported by the Science and Technology Research Project of the Education Department of Jilin Province for the 13th Five-year Plan of China (No.: 2016-379) and (No.: 2016-375).

[3] Correspongding Author: Lin Bi

## 1.Introduction

With the rapid development of the internet, various social networks expand accordingly, and massive information could follow to users. In this case, how to extract the interesting and valuable information has become critical. Thus, a variety of personalized service has come up as an emerging trend. There are multiple methods that can be used to realize the personalized topic recommendation, and the clustering algorithm is one of the typical available algorithms. In this paper, the ICIP algorithm based on clustering algorithm is applied to mass text mining, which will be of great significance. This algorithm can be used to achieve topic extraction and push forward the service of WeChat real-time hot articles.

Nan Wu, et al proposed to cluster the user situation before collaborative filtering, and combine fuzzy clustering with the collaborative filtering algorithm [1]. However, the cluster size would be too small when the similarity is comparatively high. Bowun Chen, et al introduced a trustworthy network to the recommender systems, and characterized the network with differentiated features[2]. Zhiming Xu, et al investigated an analysis technique of users and social relationship in social network [3], and classified the user attribute information. Jiajun Chen [4] studied a large-scale social network clustering algorithm, and presented a series of parallel network clustering algorithms, including the graph partition method, the module degree-based method, the spectral clustering method, etc. However, these algorithms cannot detect changes in the network structure. Sun H, et al [5] proposed an algorithm discovering dynamic network community , to analyze the change area of network topology and update dynamically by using the incremental approach. Enhong Xie divided the mixed recommendation into two ways. One is to mix content recommendation with the collaborative filtering, and the other is to incorporate both content filtering and collaborative filtering into the same framework [6]. Tsai, et al combined the K-means clustering algorithm with SOM[7]. Zongfa Cai [8] introduced the social tag to the collaborative filtering recommendation based on the graph model, and proposed a personalized recommendation algorithm based on user-object-tag tri-element relationship. In addition, the trust relationship is an important attribute of social network. Zhan Li [9] proposed to introduce the trust relationship among users into the recommendation system, which can further improve its accuracy and user satisfaction. Kehan Chen, et al [10] combined the image summary method with the content filtering recommendation algorithm, the data sparsity and cold start problem are alleviated. Hu Wu, et al [11] first performed the clustering of users and objects, and then weighted non-negative matrix decomposition in various clusters to predict the score, which could improve the accuracy of the sparse matrix. Wang S, et al [12]proposed the Convolutional Neural Network (CNN) for recommendation system, mainly to extract the hidden features in the project, so as to obtain low dimensional vectors, and combined with the hidden features for recommendations. Xu, et al [13] proposed the personalized recommendation based on DSSM， through calculating the user similarity of implicit representation to produce recommendations.

In this paper, the improved clustering based on the Isolation Point Processing (ICIP) algorithm is proposed. A clustering method aiming at WeChat hot article is designed in combination with the ICIP algorithm with text processing and analysis, which can realize the personalized recommendation. The basic principle and implementation steps of this method are elaborated in details, and the proposed algorithm is tested by using the multiple actual datasets.

## 2.WeChat Articles Recommendation Algorithm based on Clustering Algorithm

Clustering algorithm belongs to a type of unsupervised learning. It is used to classify the data whose attributes are unordered, unmatched and massive, that is to say, the system doesn't need to provide additional training set to model training and machine learning in the whole process for the purpose of overcoming the main disadvantages of K-means, which is the typical clustering algorithm. When calculating the mean of clustering, noise data or isolated points will affect the mean value and even clustering results. The processing for isolated points is included in the clustering process while using ICIP algorithm.

Outliers are different from the ordinary data in the datasets, which are the background noise in the process of clustering. According to the experiment, there exists the fact that the data are inconsistent with others in the model, which can always affect clustering efficiency. When original data includes the outlier, it is recognized as separate clusters in the clustering results, with the outcome that the remaining valuable corpus is regarded as the same cluster, which cannot be distinguished. In the process of clustering, it is necessary to identify and eliminate a large number of isolated points and then cluster the data of remaining valuable corpus. Eliminating the isolated data not only improves the clustering results, but also shortens the process of clustering.

The steps are as follows:

(1) Use n different characteristics,, to identify the whole text , and calculate weight of each feature in text d. d is abstracted to the weight of each feature in the n dimensional space as a component of the vector.

(2) The similarity among data objects is measured by the distance for each object. This algorithm uses the Euclidean distance to measure the similarity of data objects. Set algorithm through iteration and a total of k clusters are generated. The average means of all the samples in the Ci are taken. The clustering centre of Ci is set to c1, and Ci has j samples. The distance between the samples di and ci is set as distance. Set as the similarity threshold, where represents the similarity coefficient.

(3) Select k points randomly as the initial cluster centre. Calculate the Euclidean metric distance between each sample and the initial clustering centre, and allocate the samples to the nearest neighbour class according to the minimum distance principle. K clustering can be obtained.

(4) Calculate the mean values of all the samples in each Ci, and select the sample set as. Calculate the mean values of sample as new means, which could be considered as the new clustering centre of Ci.

(5) Repeat steps 2-5 until the cluster centre no longer changes.

(6) The algorithm ends and k clustering can be obtained.

## 3.WeChat Articles Recommendation Method based on the ICIP

WeChat hot articles extraction and push based on the proposed method can be achieved via the following steps.

(1) Data preparation: standardize data characteristics and reduce dimensions of the data.

In this paper, WeChat Application Programming Interface (API) is used to collect hot articles with aggregate data (https://www.juhe.cn/ ). By invoking the API, the data in JSON or XML format can be obtained to support the GET request and POST request of the HTTP protocol. After getting the required expected articles, the content of the articles is will stopped word processing, in which Chinese word segmentation is needed.

(2) Feature selection and extraction: tThe most valuable features are selected from all the features offor the sample and stored in the vector, and then new onesfeatures are formed by transforming the selected features. Based on the segmentation results, a certain weight is given for each word using by the calculation method of the weight of TF-IDF, as a measure of the lexical features of the foundation. Then the “article id-feature matrix” is built, taking each element in the matrix as the feature correspondence feature value. The two dimensional matrix is as the result of quantitative representation of text, and will become the basis for subsequent text similarity calculation and clustering.

(3) Text similarity calculation: uuse Euclidean distance to calculate the similarity of two-text vectors "article id-feature word". In the two-dimensional matrix, each line is a text eigenvector, so the process of calculating the text similarity is transformed into the process of calculating the distance between any two of the matrix row vectors. Assign the sample to the nearest neighbor class based on the principle of minimum distance.

(4) Clustering by ICIP: sselect a certain distance function (or construct a new distance function) that can suit the type of feature type to measure the similarity of data objects, and then perform clustering or grouping. The clustering results are summarized as a collection of multiple articles cantered on k cluster centre articles.

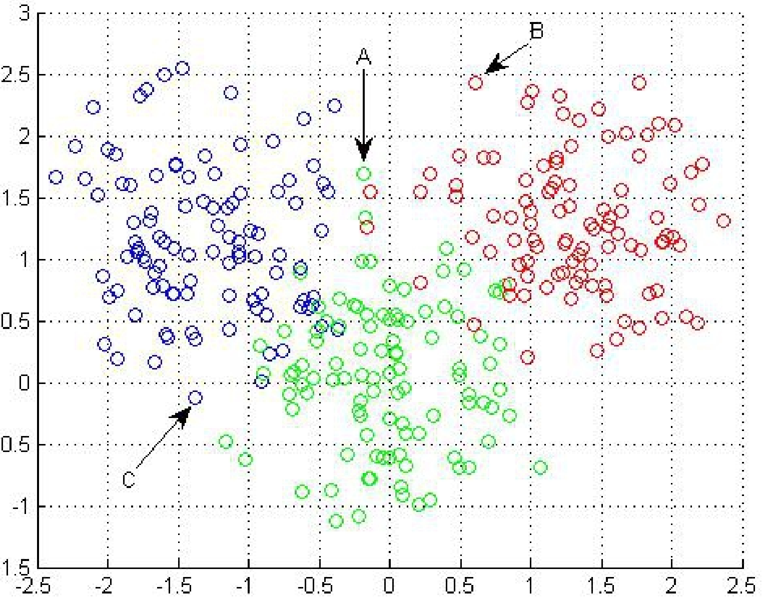


**Figure 1:** An iterative process performed by the ICIP algorithm

Figure 1 shows an iterative process performed by the ICIP algorithm.The points marked in red, green and blue represent three types of data samples, respectively. Among them, the points A, B and C indicated by the arrows are all isolated points. If the mean of all samples in each class is used in this iteration, then the clustering accuracy will be severely degraded by isolated points like A, B, C.

The implementation process of ICIP algorithm is as shown in the following pseudo code:

```
tablekmeans_parttwo(docfeaturetable,k)
1)  lastcluster ← initclusters  presentcluster ← {}  clusterresult ← initclusters
2)  while (lastcluster ≠ clusterresult)// Clustering results not convergent iteration
```

```
  new centerlist ← {}// Aalgorithm generates a collection of new clustering centres each time iterating
  for each element in last cluster
      new documents ← {}
      means ← near mean (docfeaturetable, element)// Call near mean algorithm, the most close to the sample mean vector
      for i←0 to length(initcluster)
          if(distance(element, i) ≥ threshold// Screening and clustering centre distance of similarity threshold or essay collections
              add each element to new documents
                  calculate mean vector of new documents// Calculate the mean vector collection selected articles
              insert each element of mean vector to new centre list in order// The mean vector of each element as a new clustering centre
    last cluster ← cluster result// To regenerate the clustering results
return cluster result
```

（5）Filter and recommend all types of cluster words

According to the clustering results, the collected articles have been clustered into k clusters on the basis of calculated text similarity. In order to visualize the subject of each type of article, it is necessary to filter out several characteristic words with the largest eigenvalue in all the articles in each kind of article. These feature words can provide users a reference to browse WeChat text, and help users know which kind of article is being read with high frequencyreading a higher amount so as to achieve attract the user's' interests in the subject of recommendation.

## 4.Algorithm Testing

A total of 660 articles were collected by invoking aggregated data API access to WeChat hotspots in this testing at 10 a.m. on October 11, 2017. These articles can be divided into three categories, including literature, technology, and fashion. In this paper, the traditional k-means algorithm and the ICIP algorithm are used to cluster the above articles respectively. The clustering accuracy of the two algorithms is 75.76% and 84.85%, respectively.

Figure 2 shows the relationship between k and the running time of the algorithm when the number of iterations t and the total sample data n are constant (n = 660). While tTaking into account the practical significance of the algorithm, the minimum value of k is taken as 3 and the maximum value is taken as 10. It can be seen from the Figure 2 that, when the number of samples is small, the k value increases and the running time of the algorithm does not change significantly, but with the longer the time spent toon improvinge the overall algorithm, which is consistent with the theoretical analysis. ThenTo follow that, the ICIP algorithm iswas used to keep k = 3 constant, which can increase the number of samples, and explore the relationship between the number of samples n and the running time of the algorithm. The Matlab Curve Fitting Tool was is used applied to fit the data points into a straight line. The goodness of fit (R-Square) value was is 0.9953, indicating that the model fitted well fits well to the data. The time complexity of the algorithm is linearly increased with the data size, as shown in Figure 3.

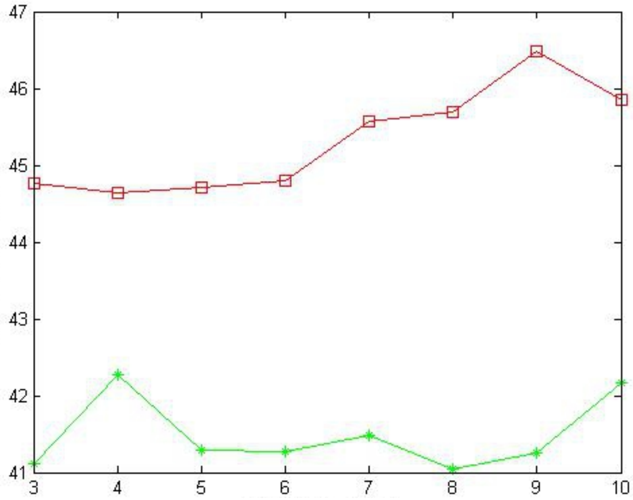
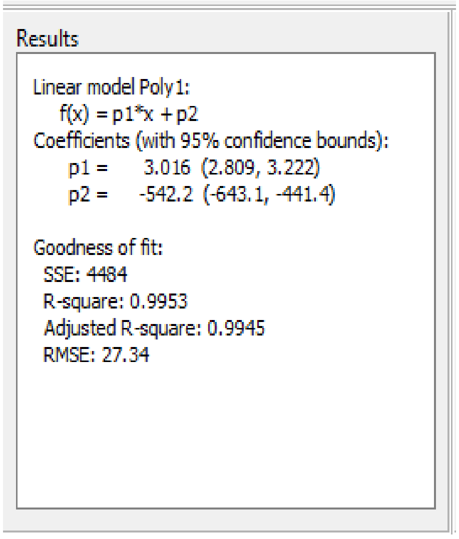


**Figure 2:** Comparison between the Traditional K-means Algorithm and ICIP on the Running run Time



```
Results

Linear model Poly1:
    f(x) = p1*x + p2
Coefficients (with 95% confidence bounds):
      p1 =      3.016  (2.809, 3.222)
      p2 =     -542.2  (-643.1, -441.4)

Goodness of fit:
  SSE: 4484
  R-square: 0.9953
  Adjusted R-square: 0.9945
  RMSE: 27.34
```

**Figure 3:** Fitting Analysis Results

## 5.Conclusions

Social networking platform can be used to provide users with a lot of information. A type of effectively personalized recommendation algorithm is proposed in this paper, which could achieve the theme recommendation based on the clustering. The ICIP algorithm steps have been introduced and also applied to WeChat hot article WeChat push service. The process for a large number of text information is demonstrated by using the ICIP algorithm process, including the process of pretreatment, feature extraction, acquiring different themes based on the clustering. In addition, WeChat articles were are used to test this method, and the experimental results show that the proposed method has better performance and accuracy, compared with other common clustering algorithms.

## References

[1]N. Wu: *A personalized recommendation algorithm based on clustering the user situation.* The Computer Technology and Development, Vol. 24(2014), p.106-109

[2]B.W. Chen: *Trust fusion tags spread and diffusion of personalized recommendation algorithm*. Computer Engineering, Vol. 12(2014), p. 33-38

[3]Z.M. Xu: Weibo user similarity measure and application. Journal of Computer, Vol. 37(2014), p. 207-218

[4]J.J. Chen: *Large-scale social networks based on structure similarity clustering algorithm*. Nankai University (2013)

[5]H. Sun, J. Huang, X. Zhang, et al: IncOrder: *Incremental density-based community detection in dynamic networks.* Knowledge-Based Systems, Vol. 72(2014), p.1-12

[6]H. E. Xie: *E-commerce recommendation technology research based on the pattern of LBS*. Beijing University of Technology (2012).

[7]C. F. Tsai, C. Hung: *Cluster ensembles in collaborative filtering recommendation.* Applied Soft Computing, Vol. 12 (2012), p.1417-1425

[8]Z. F. Cai: *Based on the user-product-tag ternary relation of personalized recommendation system research.* Value Engineering, Vol. 31(2012), p. 234-235

[9]Z. Li: *Collaborative filtering recommendation method based on social trust network research.* Dalian University of Technology (2013)

[10]J. Wu: *Heterogeneous social networks based on user clustering recommendation algorithm.* Journal of Computer, Vol. 2 (2013), p. 349-359

[11]H. Wu: *Two phase combined clustering collaborative filtering algorithm.* Journal of Software, Vol.5 (2010), p.1042-1054

[12]Wang S, Wang Y, Tang J, et al. *What your images reveal: Exploiting visual contents for point-of-interest recommendation*[C]//Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017: 391-400.

[13]Xu Z, Chen C, Lukasiewicz T, et al. *Tag-aware personalized recommendation using a deep-semantic similarity model with negative sampling*[C]//Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. ACM, 2016: 1921-1924.