APPLICATION OF DATA MINING ALGORITHM IN INTELLIGENCE ANALYSIS OF ENTERPRISE ECONOMIC INTELLIGENCE

G.E. WEI†, L. GAO‡ and F. SHI†‡

† *School of resources and environment, Hunan normal university, Hunan, China*.
‡ *School of Banking and Finance, University of International Business and Economics, Beijing, China.*
†‡ *School of Trade and Economics, University of International Business and Economics Beijing, China.*
bonnieHgTo@yahoo.com

***Abstract*— With the continuous development and application of high-speed information technology such as the Internet, the acquisition and utilization of economic intelligence has an important impact on the operation of the national economy and the operation of enterprises. Based on the detailed analysis of data mining algorithms, this paper constructs a user classification model based on clustering algorithm and a user interest feature extraction model based on UR-LDA, and uses the improved K-means algorithm in an unsupervised manner. User clustering was carried out, and data mining experiments were conducted on users of Sina Weibo. The experimental results show that the user data extracted from the interest feature topic is clustered by the improved K-means, and six similar user clusters are obtained. The better clustering results are obtained, which indicates that the classification model constructed in this paper is effective.**

***Keywords*— data mining; corporate economic intelligence; intelligence analysis.**

I. INTRODUCTION

With the continuous development of the diversification trend of the national economic system and the intensification of market competition of the enterprises, the acquisition and use of economic information have an important impact on the operation of the national economy and the operation of enterprises (Opait *et al*., 2016). With the sustainable development and application of high-speed information technologies such as the Internet, national governments have implemented macroeconomic regulation and control of the national economy, industrial restructuring, and the efficiency which companies participate in market competition is improving continuously.

Major changes have happened in the structure, management process, and operating efficiency of government and corporate economic information systems (Kim *et al*., 2016).

It is certain that China has significantly evolved in construction and operation of economic intelligence and information systems (Reinmoeller *et al.*, 2016). Notwithstanding, many of its enterprises still being in numerous problems concerning to economic intelligence, that is to say with its collection, organization, and analysis of information, under the background of the continuous development of economic globalization (Chevallier *et al.*, 2016).

The construction and operation of the economic intelligence analysis system is characterized by a combination of high systematicness, professionalism, and comprehensiveness (Opait *et al*., 2016). Therefore, this paper will use the Internet technology to study the construction of corporate economic intelligence analysis system under the data mining algorithm based on the business process of the economic system.

II. STATE OF THE ART

The research of information economics in China originated in the early 1980s. Under the macroscopic background of the Chinese government's opening-door policy, international economic operation and transnational economic cooperation must rely on the analysis and study of economic intelligence (Bulger, 2016). In the face of analysis and extraction of large amounts of data which increases continuously, in order to meet the requirements of the data age, the concept of data mining was proposed (Wang, 2016).

The competitive intelligence consciousness of American enterprises is strong, and the top managers of many large companies attach great importance to competitive intelligence. Data mining includes knowledge in many kinds of subject such as machine learning and computer technology. It also includes technology about database and mathematical statistics (Cekuls, 2015). It can analyze and extract potential laws and patterns from massive, irregular, and noisy data (Wang J. *et al.*, 2018).

After having elapsed more than 30 years of development, data mining has gradually achieved the improvement of theory, the maturity of the algorithm, and gradually realized the practical application of the specific field. For the algorithm, data mining includes: association analysis, classification prediction and cluster analysis, and so on (Brown *et al.*, 2018). All of those are suitable for different mining algorithms for different data sources in different situations. Currently, data mining has been applied in various industries, such as the retail industry (Wal-Mart's famous "beer and diapers" case), e-commerce, logistics, telecommunications, finance and other data-intensive

industries. If these industries have accumulated data, they can use data mining to find the potential value in data and it has a wide range of applications.

Furthermore, its application research in specific fields is constantly being proposed and improved, and has great development prospects (Thomas *et al.* 2016). With the deepening of specialization and the further subdivision of competitive intelligence market, competitive intelligence providers focusing on one industry or region emerge (Falco *et al*., 2014).

III. METHODOLOGY

A. Construction of Corporate Economic Intelligence Information System Based on Clustering Algorithm

Clustering, as the name suggests, aggregates objects with similarities in a data set to achieve data set classification. Each of these types is a collection of data objects (Sheng *et al*., 2014). Data objects in a collection have a high degree of similarity, while different types of data objects have a lower similarity. Second, the feature of clustering as non-supervised learning is that it can automatically divide data objects without manual processing.

The construction of corporate economic intelligence information system based on clustering algorithm, through the analysis of the structure and operation logic of the economic system, we can see that economic intelligence information has an important impact on the stable and effective operation of each economic unit.

The establishment of economic information systems requires the analysis of economic information needs, the plans of the establishment of economic intelligence information system construction, the decomposition of economic information structures, the determination of operation processes in economic intelligence system, the definition of economic intelligence information system functions, and the collection and analysis methods of economic intelligence information.

Based on the operating mode of the economic intelligence information system, it actively promotes the realization of the monitoring of economic intelligence information, the early warning of changes in economic systems and economic units, the evaluation and adjustment of economic intelligence information systems, the analysis of economic intelligence information, and the operating strategies of economic systems and economic units and economic intelligence and information security.

At present, researchers in various fields have proposed many clustering methods and algorithms, but it is difficult to propose an effective clustering method suitable for various fields.

However, clustering methods still have a broad classification: partition-based methods, hierarchy-based methods, density-based methods, and grid-based methods, each of which has its typical, commonly and used algorithms. A detailed vision of the structure is shown in Fig. 1:

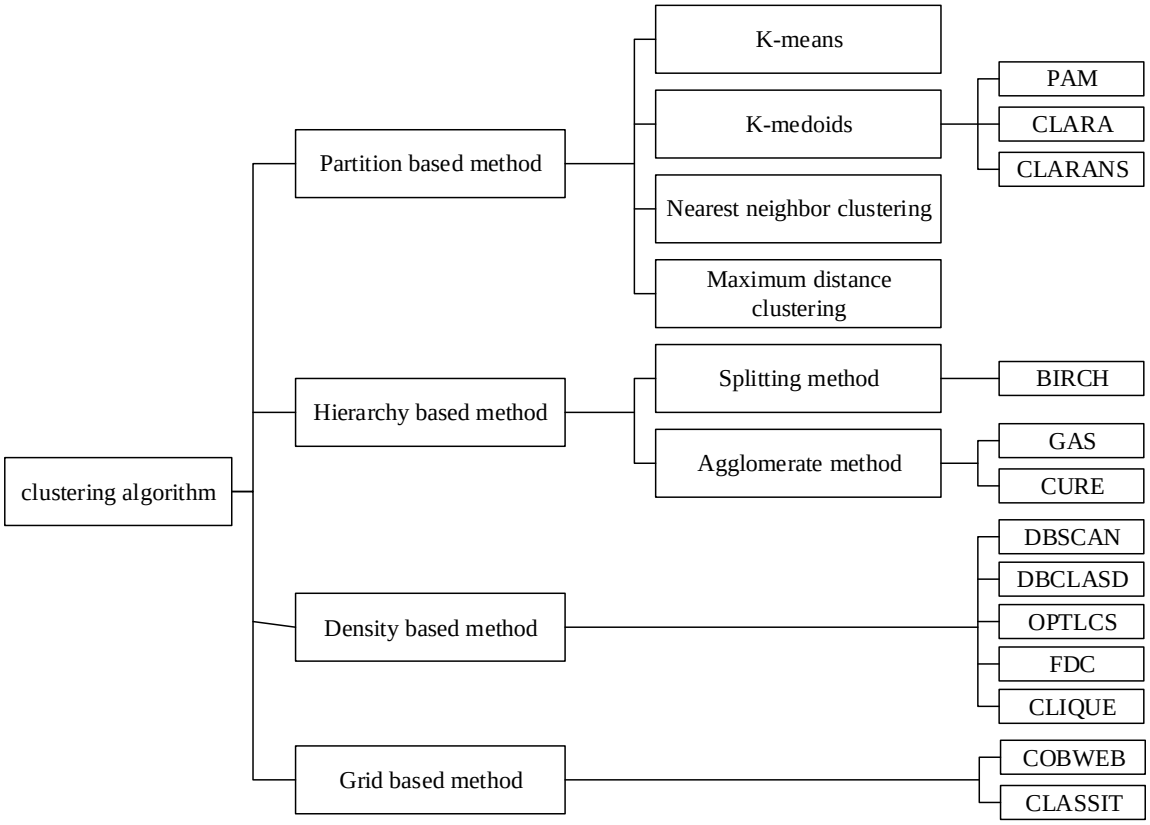

Figure 1. Structure diagram of clustering method.

B. K-means Algorithm and Optimization Based on Data Mining

In 1967, MacQueen proposed the K-means algorithm and was chosen as the top ten classical algorithms. The algorithm is used by a large number of academic researches and a lot of industry analysis because of its convenient application and simple implementation. The basic idea of the K-means algorithm is to set up k classes, divide all data objects into the required number of classes through continuous iteration, calculate the distance from each data point to its center, and ensure that the distance is the minimum value. Then the global optimal solution can be realized.

In the K-means algorithm, the first is the determination of the k initial cluster centers. Through continuous algorithm iteration, the cluster center is continuously optimized, and the data objects are adjusted back and forth in different clusters, and the optimal division of the data set is finally achieved. The flow of the algorithm is shown in Fig. 2.

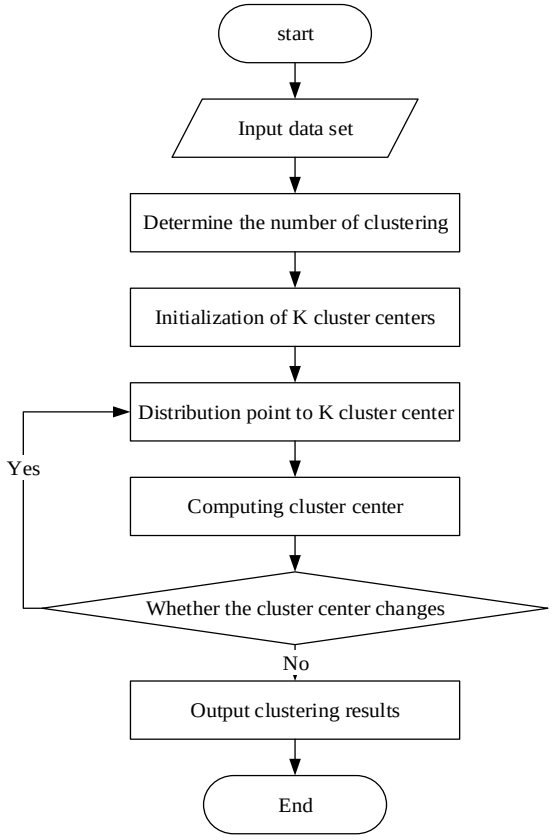

Figure 2. Flow chart of the K-means algorithm.

The selection of the initialization center is achieved by using the cutting distance in the sample data set. It will find the initial clustering center of data by equal distances mainly through the calculation of the two data points farthest from each other because the difference between the two sample points is the largest, and almost all the points of the data set can be contained.

First, the distance between all points in the data set is calculated by using the Euclidean distance formula (Eq. 1), and the two points farthest away are found. Among them, any two points in the data set is called υ and ω respectively, and the dimension of the data point is i, and the distance between the two points is d, as seen in Equation 1:

$$d(v-w)=\sqrt{\sum_{i=1}^{n}(v_i-w_i)}^{2}\ i=1,\ldots,n \tag{1}$$

Then, the cutting distance of each dimension is calculated. The formula is shown in Eq. 2, which is mainly calculated by the known number of clusters k.

$$d_{cut}^{i}=\frac{v_i-w_i}{k-1}\ i=1,\ldots,n,k>2 \tag{2}$$

in Eq. 2: d_{cut}^{i} t is the cutting distance in this dimension, i is the dimension of the sample point, and k is the number of clusters. If k=2, no calculation is needed by the system, and the initial cluster center is the two points farthest away. If k>2, then the initial cluster center needs to be calculated by the cutting distance. Since each dimension needs to be calculated independently, it is possible to calculate the cutting distance in each dimension, and the cutting distance in each dimension is independent. The formula is shown in Eq. 3:

$$z_j^i=z_{j-1}^i+d_{cut}^i \quad j=2,\ldots,k-1 \tag{3}$$

In Eq. 3, the z represents the cluster center. For example, v and ω are two three-dimensional sample points.

Determining the similarity between users is an important step that needs to be carried out before the user clusters. Then, the similarity calculation between users seems to be essential.

At present, there are many methods for calculating similarity between vectors in previous research. In this paper, the similarity of interest features between two users will be calculated by the cosine similarity. The cosine similarity can be calculated for multi-dimensional vectors, which is consistent with the multi-dimensional features of social network users. In n-dimensional space, the similarity between two users is evaluated by the cosine of the angle between two vectors. The corner size represents the degree of similarity between the interest features of two users. The smaller the angle is, the higher the degree of similarity of the user's interests is, otherwise, the lower the degree of similarity is.

The K-means algorithm firstly determines the number of clusters k in the process of implementation. However, for Sina Weibo user data lacking clustering experience, it is not possible to determine the most appropriate number of classes, and the optimal k value needs to cluster for many times to see the effect of clustering. In order to face this problem, the clustering number k is evaluated by the cluster validity function DBI index to determine the optimal clustering number k.

The DBI (Davidson Fortary Index) is an evaluation index for non-ambiguous clusters based on the principle of geometric operations. Its evaluation is based on the degree of closeness of data points and the degree of dispersion between different clusters in the same cluster. That is, the similarity within a class is positively correlated with the size of the index, and the similarity between classes is negatively correlated with the size of the index. When the distance between the data points in the cluster is smaller and the distance between the clusters is larger, the DBI value is smaller, which means that the differences between the clusters are large and the differences within the clusters are very small. Therefore, the smaller the DBI index for different clustering numbers is, the closer the clustering number is to the number of clusters in the dataset itself.

The methods for collecting economic information mainly include:

(1) Collecting intelligence information for various types of books, magazines, newspapers, and special research reports related to the operation of economic units;

(2) Various types of Internet information platform related to the operation of economic units;

(3) Collecting intelligence information including government websites, commercial portals, professional and technical websites, and key corporate websites;

(4) Obtaining economic intelligence information through economic surveys, key surveys, and sample surveys.

In summary, the construction of economic intelligence analysis system is a complex system engineering. Requirements analysis, planning and formulation, system structure, operation flow, main functions, and related methods are the core elements of the design and construction of economic intelligence information systems. In the analysis of user interest characteristics, if there are certain linear relationships between the two analysis factors, then the theme of two analysis factors will have very similarities or connections. Therefore, whether there is a linear relationship between the two analysis factors is very important for the study. If there is a large linear relationship, such a user's analysis is meaningless. Therefore, a correlation test is required. Before the user analysis is performed, this paper uses Pearson correlation detection to calculate the correlation coefficient of each analysis factor.

IV. ANALYSIS AND DISCUSSION OF RESULTS

The parameter of K-means cluster number was set during the experiment. Since the interest of the user in the Weibo has been classified into 10 categories, the number of user clusters is $k_{\max} = 10$. The K starts from 2 and increases by 1 per cycle until k=10. Through the execution of the clustering algorithm, the initial center is generated randomly, and the value of the current evaluation index is calculated at the end of different k-value clusters.

The experimental environment is MatlabR2011. The result is shown in Fig. 3:

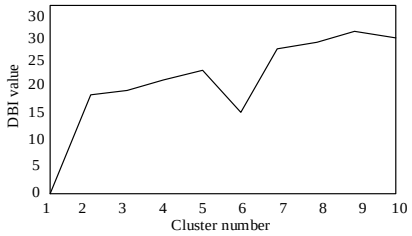

Figure 3. The DBI value of the number of clusters.

From the experimental results, it can be concluded that when k=6, the DBI value is 16, which is the number of clusters that meet the requirements. Therefore, when using k-means for clustering, microblog users will be clustered according to 6 categories. The improved K-means algorithm mentioned above cluster the microblog user data. 5000 topics of interest of Weibo users in Chapter 3 are taken as classified data sets, and finally users are divided into 6 categories. The data set is shown in Table 1.

Table 1. Cluster results dataset.

Classification	Text quantity	Classification	Text quantity
Data set A	520	Data set D	1110
Data set B	715	Data set E	185
Data set C	640	Data set F	1830

In addition, the distance between cluster centers is calculated, as shown in Table 2.

Table 2. Class center distance.

Clustering	A	B	C	D	E	F
A		0.313	0.421	0.334	0.264	0.380
B	0.313		0.580	0.465	0.386	0.556
C	0.421	0.580		0.492	0.457	0.541
D	0.334	0.465	0.492		0.341	0.384
E	0.264	0.386	0.457	0.341		0.512
F	0.380	0.556	0.541	0.384	0.512	

From the data in Table 2, the six final cluster centers have significant differences of distance, and the distance between center 2 and center 3 reaches 0.580, which is the center of the greatest difference, and the distance between center 1 and center 5 is the smallest, which is 0.264. Based on the results of user classification, the improved clustering algorithm is evaluated. In this paper, clustering results evaluation index F are used to evaluate the result. The indicators of evaluation F mainly include Recall and Precision. The effect of clustering results is evaluated through recall and precision. It cannot only accurately measure clustering results, but also can intuitive clustering effects easily.

Experiments show that the improved k-means method enhances the recall rate, precision rate, and F value compared with the traditional k-means algorithm. It is proved that the improved K-means algorithm is feasible and can realize the cluster of Weibo users better. For observation and analysis in an easy way, we made

the average results of the clustering results and the values of the various attributes of the generated 6 classes, as shown in Table 3.

By discretizing the cluster result data, the user's interest preference is more intuitive. Through the value of the range of interest variables, the definition of weak interest is [0-0.1], the moderate interest is [0.1-0.2], and the strong interest is [0.3-1].

From Table 3, we can see that there are six types of users who are interested in "life style" and "reading", showing that people should generally have these two types of interests. The six categories of ordinary users get interested in "life style" and "reading" which mainly include fashion brands (bloggers), shopping, and tidal network red microblogs, covering popular fashion trends.

"Reading" mainly includes literary appreciation, life encyclopedia, chicken soup for the soul, microblogs of various routines about skill, covering the needs of reading, literary resonance, and practical knowledge. These two types of interests are closely related to people's lives. The moderate interest displayed by users indicates that Weibo users pay attention to the life they are living at, and at the same time, they continuously improve their own quality and acquire knowledge. Therefore, in the process of recommending push, in addition to the "possible person", we can recommend some microblog users that are representative of "life style" and "reading". With the exception of these two types of interest variables, each major category performs differently in other categories of interest variables and does not have universal applicability.

Table 3. Average of interest variables in 6 users' categories.

Category	A	B	C	D	E	F
Life fashion	0.15	0.14	0.17	0.47	0.13	0.12
Gourmet Entertainment	0.05	0.07	0.13	0.13	0.04	0.04
Photography tourism	0.11	0.08	0.11	0.06	0.07	0.07
Sports	0.32	0.00	0.03	0.01	0.02	0.02
Film	0.03	0.07	0.12	0.03	0.13	0.04
Music	0.07	0.02	0.02	0.05	0.07	0.02
Anime game	0.12	0.02	0.03	0.02	0.05	0.01
Reading	0.10	0.43	0.10	0.10	0.11	0.15
Industry current affairs	0.01	0.13	0.11	0.09	0.05	0.45
Digital electronic	0.01	0.01	0.05	0.03	0.37	0.08
Sample size	520	715	640	1110	185	1830

In recent years, the rapid development of Weibo has quickly occupied our lives. For the maintenance of Weibo users, on one hand, it is necessary to provide them with information that meets their interests and enhances the user's viscosity continuously; on the other hand, Weibo has a powerful user base. For the development of company's e-commerce, it is necessary to grasp the characteristics of the user base. The following discusses strategic recommendations in both information recommendation and marketing promotion.

For the six types of users detailed above, according to the characteristics of each type of users, different suggestions and considerations are proposed:

The A type of users should be sports enthusiasts who are mainly students. They have plenty of free time and most of them are fans of sports, movies and telegrams. You can recommend sport events and movies and TV shows for them. As sports enthusiasts, they are also very interested in some fitness activities and sports celebrities. This aspect should also be considered.

In the case of type B users, it can be highlighted that they have relatively large needs in the information of reading and acquiring practical applications, and they have higher demand in the need of information. We should recommend relevant professional data and magazines. Users should first order specific fields (history, economy, literature) and then according to their field, the influential Weibo is recommended.

For type C users, the interests and preferences are basically the same and there are no prominent points of interest. Therefore, they are recommended to read a wide range of information. It does not stick to a certain type of information, and we can recommend some hot microblogs in daily life, such as TV propaganda, news and pictures.

For D type users, they like noble lifestyles and material entertainment, and pursue innovative material life. For them, the microblogging platform should recommend senior fashion to them, such as fashion bloggers, trend collocations, pop stars, etc. In addition, some local foods, such as local food, snacks and other things can also be recommended.

For E type users, as digital electronic enthusiasts, they may engage in work related with Internet and study for a long time. They want to know the information about industry, and we should recommend some real-time industry trends in response to their needs. Other new electronic products must be recommended, such as smart bracelets, robots, computers, and another practical microblog information should also be considered.

For F type users, they are professional in industry. From the aspect of referrals, we should recommend industry trends and practical information that meets their needs, such as automobiles, real estate, and finance.

The user selects the industry sector and the recommended industry should be valuable microblogging of the industry. At the same time, we should also consider recommending some useful microblog information to provide users with work, study, and life help, such as studying abroad, vocational training, and microblogs of famous teachers.

V. CONCLUSIONS

In this paper, a user interest feature extraction model based on UR-LDA is proposed by using improved K-means algorithm. User clustering in unsupervised way is used to study the construction of enterprise economic intelligence analysis system under data mining algorithm.

The results show that by integrating the construction of economic intelligence analysis system, evaluating the overall objectives of economic intelligence information system construction, operation and management, phased objectives and human resource demand trends, we can provide a guarantee basis for promoting the construction of economic intelligence information system. User data extracted from interest feature topics are clustered by improved K-means. Six similar user clusters are obtained and good clustering results are obtained.

REFERENCES

Brown, T., S. Du, H. Eruslu, and F. J. Sayas, "Analysis of models for viscoelastic wave propagation," *arXiv preprint*, 1802.00825 (2018).

Bulger, N.J., "The evolving role of intelligence: migrating from traditional competitive intelligence to integrated intelligence," *The International Journal of Intelligence, Security and Public Affairs*, **18 (1)**, 57-84 (2016).

Cekuls, A. "Leadership values in transformation of organizational culture to implement competitive intelligence management: the trust building through organizational culture", *European Integration Studies*, **23 (9)**, 244-256 (2015).

Chevallier, C., Z. Laarraf, J.S. Lacam, A. Miloudi, and D. Salvetat, "Competitive intelligence, knowledge management and coopetition: The case of European high-technology firms," *Business Process Management Journal*, **22(6)**, 1192-1211 (2016).

Falco, L., A. Martínez, M.P Di Nanno, H. Thomas, and G. Curutchet, "Study of a pilot plant for the recovery of metals from spent alkaline and zinc-carbon batteries with biological sulphuric acid and polythionate production," *Latin American Applied Research*, **44(2)**, 123-129 (2014).

Kim, Y., R. Dwivedi, J. Zhang, and S.R. Jeong, "Competitive intelligence in social media twitter: iPhone 6 vs. Galaxy S5," *Online Information Review*, **40(1)**, 42-61 (2016).

Opait, G., G. Bleoju, R. Nistor and A Capatina, "The influences of competitive intelligence budgets on informational energy dynamics," *Journal of Business Research*, **69(5)**, 1682-1689 (2016).

Reinmoeller, P. and S. Ansari, "The persistence of a stigmatized practice: a study of competitive intelligence," *British Journal of Management*, **27(1)**, 116-142 (2016).

Sheng, Z., Y. Wang, P. Wan and Y. Li, "Ultrasound-assisted extraction of total flavonoids from leaves of Syringa Oblata Lindl," *Latin American Applied Research*, **44(2)**, 131-135 (2014).

Thomas, J. and R.T. Princy, "Human heart disease prediction system using data mining techniques", *Circuit, Power and Computing Technologies, IEEE*, **1(1)**, 1-5 (2016).

Wang, J. T. Lingyu, Z. Xianyong and L. Yuyan, "Three-way weighted combination-entropies based on three-layer granular structures", *Applied Mathematics and Nonlinear Sciences*, **2(2)**, 329-340 (2018).

Wang, Y.H., "Identifying competitive intelligence of collaborative intellectual property alliances: analytic platform and case studies," *Information Systems and E-bus Management*, **14(3)**, 491-505 (2016).

Received: December 15th 2017
Accepted: June 30th 2018
Recommended by Guest Editor
Juan Luis García Guirao