Research in Astronomy and Astrophysics, 22:065015 (13pp), 2022 June https://doi.org/10.1088/1674-4527/ac693b

106 New Emission-line Galaxies and 29 New Galactic H II Regions are Identified with Spectra in the Unknown Data Set of LAMOST DR7

Yan Lu[1,2,3], A-Li Luo[1,2,3], Li-Li Wang[2], You-Fen Wang[1], Yin-Bi Li[1], Jin-Shu Han[2], Li Qin[2], Yan-Ke Tang[4], Bo Qiu[5], Shuo Zhang[6,7], Jian-Nan Zhang[1], and Yong-Heng Zhao[1,3]
[1] CAS Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China; lal@nao.cas.cn
[2] College of Computer and Information Engineering & Institute for Astronomical Science, Dezhou University, Dezhou 253023, China
[3] University of Chinese Academy of Science, Beijing 100049, China
[4] College of Physics and Electronic Information, Dezhou University, Dezhou 253023, China
[5] School of Electronic Information Engineering, Hebei University of Technology, Tianjin 300401, China
[6] Department of Astronomy, School of Physics, Peking University, Beijing 100871, China
[7] Kavli institute of Astronomy and Astrophysics, Peking University, Beijing 100871, China
Received 2022 February 3; revised 2022 April 6; accepted 2022 April 20; published 2022 May 26

Abstract

This work is to retrieve emission-line spectra from the "Unknown" data set in LAMOST DR7 V1.2, most of which are low signal-to-noise ratios spectra. In the work, we perform emission line search and redshift calculations on the Unknown data set to get possible emission line galaxy spectra. Taking the galaxy spectra released by LAMOST as templates, the Product Quantization (PQ) based approximate nearest neighbor (ANN) search is used to retrieve the nearest neighbors of each spectrum. We keep the spectra for which the calculated redshift and the published redshift of the template meet the threshold, and 16,188 spectra with emission lines are obtained from the LAMOST DR7 Unknown data set. After visual inspection of spectra 10,266 spectra are left, in which 5828 spectra are identified as emission-line galaxies, 1782 spectra show ionization nebula features, and other 2656 are not clearly classified. Among 5828 spectra, 5720 can be found in Strasbourg astronomical Data Center catalog, Sloan Digital Sky Survey catalog, or NASA/IPAC Extragalactic Database catalog. The 108 spectra (corresponding to 106 unique coordinates of R.A. and decl.) which have no record in these three databases are new discoveries; for the 1782 spectra showing ionization nebula features, most of them have radial velocities less than 150 kilometers per second. We check them with the latest version of WISE H II catalog V2.0 (short for HIICat V2) and 985 out of the 1782 spectra belong to 72 H II regions. Of these H II regions, 43 were previously identified while the other 29 are newly identified in this work including 797 newly observed spectra. Besides, there are still 2656 spectra that cannot be clearly classified although they have obvious emission lines and with small redshift. Finally, 106 new emission-line galaxies and 29 new Galactic H II regions are identified, and we conclude that the ANN method sped up by the PQ algorithm is efficient in solving the problem of pairing spectra with massive data set to figure out their classes. We present our result at the link http://paperdata.china-vo.org/LY_paper/Work2/pressWork2_last.zip.

Key words: techniques: spectroscopic – Galaxies – (ISM:) HII regions

1. Introduction

As we know, an important fraction of galaxies has narrow emission lines in their spectra such as [S II]λ6718,6732,Hα and [N II]λ6585,6548, Hβ, [O III]λ5007,4959, [O II]λ3727, [O II] λ6300 etc. which refer to the vacuum wavelengths corresponding to the published data of LAMOST (Veilleux & Osterbrock 1987). Galaxies are mainly divided into star-forming galaxies, Seyfert galaxies, LINER galaxies, starburst galaxies (Kennicutt 1992; Kauffmann et al. 2003; Brinchmann et al. 2004; Yip et al. 2004). There are some important archived databases of galaxies, such as Strasbourg astronomical Data Center (CDS) (Ochsenbein et al. 2000) catalog, Sloan Digital Sky Survey (SDSS) (Blanton et al 2017; Eisenstein et al 2011), and the NASA/IPAC Extragalactic Database (NED) (Helou et al. 1995), etc, which include most of known galaxies. Besides these released galaxy spectra, LAMOST has also conducted a spectral survey which produced more than 200 thousand galaxy spectra.

H II regions are ionized gaseous nebula whose spectra usually have emission line features, such as hydrogen and metal elements (Baldwin et al. 1981; Anderson et al. 2011). There are many observations and studies of Galactic H II regions. Esteban et al. 2017 presented deep optical spectroscopy of eight H II regions in the anti-center direction of the Milky Way and found the absence of flattening in the radial oxygen gradient in the outer Milky Way. Anderson et al. (2014) provided a Galactic H II region catalog based on the Wide-Field Infrared Survey Explorer (WISE) project which contains about 1500 sources in the anti-center direction of the Milky Way, and the catalog has been updated by Anderson et al. (2015, 2018).

Wang et al. (2018a) spectroscopically identified 101 Galactic H II regions in the Galactic Anti-center Area using spectra from LAMOST survey.

Since the pilot survey (Luo et al. 2012), LAMOST has released over ten million low-resolution spectra in LAMOST DR7 v1.2. Besides released stars, galaxies, and QSOs, there are 492,938 spectra classified as "Unknown" by the LAMOST 1D pipeline. The pipeline based on cross-matching with templates refused to classify these "Unknown" spectra because of the limited classification capabilities and limited coverage of matching templates. However, it is still possible to classify part of them which have obvious emission lines. If we can identify the emission lines, we probably classify these spectra. If we can find that they are highly similar to the known spectra, these spectra are likely to be mis-classified emission line galaxies spectra. In this paper, we select possible spectra having emission lines out of nearly 500,000 unknown data, and search for approximate nearest neighbors for them using the published galaxy spectra as template. In order to overcome the bottleneck of such a large computational complexity, we use the product quantization (PQ) based approximate nearest neighbor search (Jégou et al. 2011) as the solution.

The PQ algorithm can quickly retrieve the nearest neighbor vector of a known query vector under limited time and space conditions. This is an excellent method for retrieving approximate nearest neighbors in large-scale data. Each vector in the database is quantified into a short PQ code, and the search is completed by querying the PQ code according to the table. In this paper, we use PQ based on approximate nearest neighbor retrieval as a tool to obtain spectra with strong emission lines in LAMOST "Unknown" data set using LAMOST published galaxies spectra as templates. Also adopted in this work is the method that combines characteristic emission lines to get redshift and other operations about the spectral data are applied. Every spectrum is segmented into several parts, and the K-means clustering is run on each segment for all samples in one run. Then we encode the cluster center to get codewords of each segment, which are operated by Cartesian product to form the whole codebook of our retrieval work. When a query occurs, the distance from different codewords can be calculated on each segment (we use Euclidean distance), thus forming a distance table only related to the number of segments and the number of cluster centers. The distances between the query and every item to be retrieved can be directly cascaded by the distances of each segment. In this way, we get a sequence of all the items to be retrieved from the smallest distance to the largest distance. We choose the target with the smallest distance as the approximate nearest neighbor of the query.

Confirmation of redshift is necessary to identify galaxies and H II objects. By detecting and fitting the emission lines, we recognize the lines and get the redshift of each object. Comparing the redshift we measured with the redshift of the approximate nearest neighbor (template), we can confirm new galaxies and candidates of the H II region. Through cross-matching with the HIICat_V2, we can identify Galactic H II regions.

The paper is organized as: In Section 2, we introduce the spectral data from LAMOST DR7 and the data operation used in our work. Section 3 describes the product quantization-based approximate nearest neighbor retrieval method and our work strategy. In Section 4 we demonstrate how to apply the ANN method to the "Unknown" data set of LAMOST DR7 and we give the result analysis. In Section 5, we give the discussion and sum up our study result.

2. Spectral Data and Data Processing

2.1. Data from LAMOST DR7 v1.2

The Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST) is a Chinese national scientific research flagship facility operated by the National Astronomical Observatories, Chinese Academy of Sciences. LAMOST is a specially designed spectroscopy survey telescope with 4000 fibers in a field of view of 20 degree^2. It is equipped with 16 spectrographs, each of which is fed by 250 fibers (Cui et al. 2012). LAMOST began a five-year regular survey in 2012 September, which includes the LAMOST ExtratraGAlactic Survey (LEGAS) and the LAMOST Experiment for Galactic Understanding and Exploration (LEGUE) (Deng et al. 2012; Zhao et al. 2012). As Luo et al. (2015) show that the raw stellar spectra are first processed by the LAMOST 2D pipeline, then LAMOST 1D pipeline performs the spectral type classification and radial velocity measurement.

Until July 2018, LAMOST has completed its pilot survey (Luo et al. 2012), which was launched in October 2011 and ended in June 2012, and the first seven years of regular survey, which was started in September 2012. In LAMOST DR7 database (Napolitano et al. 2020), there are totally 10,599,979 low resolution spectra published, including 9,842,272 stellar spectra, 198,440 galaxy spectra, 66,329 quasar spectra and 492,938 unknown object spectra, which cover the wavelength range of 3690–9100 Å with a resolution of ∼1800 at 5500 Å. The LAMOST spectral analysis pipeline "v2.9.7" determines the classification of the published spectra with labels "STAR", "GALAXY", "QSO" or "Unknown". In this work, we use the published spectra of "GALAXY" of LAMOST DR6 as template spectra, and retrieve potential galaxy spectra that may be mislabeled in the "Unknown" data set of LAMOST DR7.

2.2. Spectral Data Processing

2.2.1. Data Preprocessing

For consistency in the overall work, we only keep the spectral data in the wavelength range between 3900 and 8900 Å for each spectrum in our data set. According to Xiang et al. 2015, the typical full width at half maxima (FWHM) of

LAMOST spectra is ∼2.8 Å. We use all integer steps (values of 1 and 2) smaller than the resolution of the spectrometer to interpolate the spectrum. The spectra are interpolated within 3900–8900 Å with a step of 2 Å which has a total number of sampling points 2500, while interpolated to the same wavelength range with a step of 1 Å there are 5000 sampling points. Then all the interpolated spectra are normalized to the same scale in different dimensions to ensure the effectiveness of data processing, and the normalization formula used is: $f_i = \frac{f_i}{\sum_{j=1}^{N} f_i^2}$. The numerator represents the flux value of each interpolation sampling point, and the denominator is the sum of the squares of the flux values of all sampling points in a spectrum.

2.2.2. Searching for Strong Emission Lines

First, we use median filtering to get an approximate pseudo-continuum, which will then be subtracted from the preprocessed flux to get the spectral flux residual containing the information of corresponding wavelength range of the strong emission lines. Next, we use the sliding window method (abbreviated as SW) to get the central wavelength position of the strong emission lines and the start and end points of its wavelength range. The sliding window method refers to applying a moving window to a given array for the required operation.

2.2.3. Redshift Computing

In physics and astronomy, redshift refers to the phenomenon that the frequency of electromagnetic radiation of an object decreases. When redshift occurs, the spectral line moves toward the red end for a certain wavelength in the visible light band, and the ratio of moving wavelength of all spectral lines to their vacuum wavelength in a spectrum is unique. The redshift is represented by z, and the calculation formula of the redshift value is defined as $z = \frac{\lambda - \lambda_0}{\lambda_0}$, λ is the observed wavelength of a spectral line, and λ_0 is the rest wavelength. In our work, we check whether there are multiple emission lines (obtained using SW) that can match the prominent features in the spectrum of galaxy simultaneously, and when three lines (at least three) satisfy the criteria, we can get the redshift using the redshift formula. This method of combining characteristic emission lines used to get redshift is noted as CL. Redshift error threshold for judging whether the combined emission line matches the characteristic spectral line successfully is explained in Section 2.2.5.

2.2.4. Redshift Confirmation

Using PQ based retrieval method, we get potential candidate spectra having emission lines corresponding to the galaxy templates. We need to set a redshift matching threshold as the criteria for judging whether the calculated redshift matches to the published redshift of the template. How to set the redshift matching threshold will be introduced in Section 2.2.5. Every spectrum gets two redshift results, one is the calculated redshift obtained by the CL methods, and the other is the published redshift of the corresponding template which is called "redshift transfer method" (short as RT). The RT method can help us get redshift value of the target spectrum conveniently and quickly. If the redshift value of CL matches to the one of RT, we can prove the accuracy and reliability of our work.

2.2.5. Threshold Configuration

We make a series of threshold configurations to the operation for the spectral data. In searching for strong emission lines, the strong emission interception threshold of galaxy spectrum is set to $\frac{1}{10}$ of the overall spectral flux values from large to small sequence. The second configuration is to determine the error threshold of redshift value obtained by different characteristic emission line centers in the process of redshift calculation using the CL method. We set 0.001 as the error threshold for the six prominent emission lines: Hα, Hβ, [O III]λ5007 [O III]λ3727, [S II]λ6717, [N II]λ6585. Next is the third configuration. After using our method to get redshift value, we need to set the redshift matching threshold to filter those that can match the redshift value issued by LAMOST. Similar to the previous threshold determination process, we use 0.001 as the matching threshold for galaxy spectra.

3. Method and Work Strategy

In our work, the approximate nearest neighbor retrieval method based on product quantization is used to pick out our candidate spectra. We elaborate on this main method in this section, including the basic principles of PQ, the implementation of the PQ method in spectral data, the validity of the PQ method in spectral data retrieval and our work strategy.

3.1. Product Quantization Based Approach for Approximate Nearest Neighbor Search

3.1.1. Approximate Nearest Neighbor Search

Given a D-dimensional data set $S = \{X_n\}_{n=1}^{N}$ ($X = (x_1, x_2, \ldots x_D)$) and one query item y, with $y \in S$, the nearest neighbor search problem is to find the item $NN(X_n)$ minimizing the distance to the query item y:

$$NN(X_n) = \underset{X_n \in S}{\operatorname{argmin}} \|y - X_n\|_2 \quad (1)$$

Equation (1) can be resolved by linear searching distance calculation, whose complexity is $O(DN)$. In order to reduce the search time, the approximate nearest neighbor (ANN) search has been proposed, which aims to find the nearest neighbor with high probability, instead of probability 1. There are some popular methods belonging to ANN such as KD-tree, LSH (Locality

Sensitive Hashing) that have shown excellent performance. Another method suitable for large-scale data retrieval is the vector quantization method, and its typical representative is the product quantization method used in our work.

3.1.2. Basic Principles of PQ

The function of the PQ method is to decompose the original vector space into Cartesian products of several low dimensional vector spaces. The decomposed low dimensional vector spaces are quantized separately. Thus each vector can be represented by the combination of quantization codebooks of low dimensional vector. The specific implementation process of the PQ method can be described as follows:

The D-dimensional space is divided into M subspaces and each subspace contains D/M-dimension, then the K-means clustering method is carried out in each subspace. A given vector X is mapped as follows ($u_j(x)$ presents the j-th subvector with $j \in \{1, 2, \cdots, M\}$):

$$(x_1,\cdots,x_{D/M}),\cdots,(x_{D-D/M+1},\cdots,x_D) \rightarrow u_1(x),\cdots,u_M(x) \quad (2)$$

The subvectors are quantized separately using M distinct quantizers to get M independent sub-codebook: $C^M = \{c_k\}_{k=1}^K$. We perform clustering with the number of cluster centers K (user-defined) for each subspace of all vectors, and the j-th part of each vector is represented by the cluster center label to which it belongs, which is the sub-codeword c_k. The codebook is therefore defined as the Cartesian product:

$$C = C^1 \times \cdots \times C^M \quad (3)$$

The PQ algorithm uses the method of asymmetric distance to calculate the distance between the codebook obtained by segmentation and clustering. When a query item occurs, first we divide it into segments according to the same configuration. Then we calculate the distance between each segment of the query and the cluster centroids obtained by the foregoing process. Now we get the $M \times K$ distance table. Thus, when calculating the distance between the query item and each item in the data set, we can directly get the distance of each segment in the distance table according to the sub-codeword c_k in the sub-codebook C^M. Next we use Cartesian product method to connect the distance values of each part to get the distance between the query and each item. Instead of directly calculating the distance between two vectors, the method of calculating the representative distance using the centroid of the cluster where one of the vectors is located is the asymmetric distance method.

Using different numbers of segments and cluster centers, the PQ method can be used effectively to get the nearest neighbors. Thus, this method is helpful to select similar targets of known types from a large amount of data set.

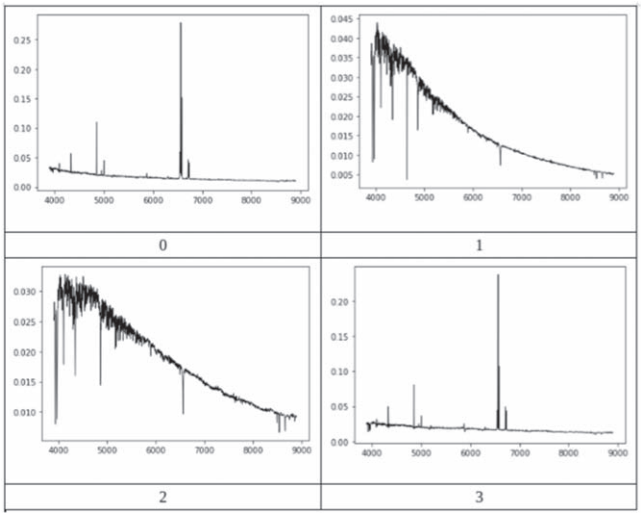

Figure 1. Examples used to illustrate the implementation process of product quantification in spectral data.

3.1.3. Implementation of PQ Method in Spectra

We take four spectra shown in Figure 1 as examples to illustrate the implementation process of product quantification in spectroscopy applications. We remove the sampling points not belong to the wavelength range between 3900 and 8900 Å in each spectrum, and we interpolate all the spectra to 3900–8900 Å to get the corresponding fluxes in step of 2 Å. And we standardize all the fluxes in a unified way and the normalization formula used is: $f_i = \frac{f_i}{\sum_{j=1}^N f_i^2}$. Then we divide each spectrum into 25 segments, with each segment spanning the wavelength range of 100 Å . In each wavelength range, we cluster the four different spectrum segments using the K-means method, and the number of cluster centers is 2. In this way, 2 cluster centers on every 25 segments can be obtained as shown in Figure 2. The numbers arranged horizontally are the subscripts of the segments (0–24), and the vertical numbers on the left are the flags of the cluster centers (0, 1). Thus, we can compress the 4*2500 original spectra as a list of

$$4^*25: \begin{bmatrix} 0&0&1&1&1&0&0&1&1&0&0&0&0&0&0&0&0&0&1&1&0&0&0&0&0 \\ 1&1&0&0&0&1&1&0&0&0&0&0&0&1&1&0&0&1&0&0&1&1&1&1&1 \\ 0&0&0&0&0&1&1&0&0&1&1&1&1&1&1&1&1&0&1&1&0&0&0&0&0 \\ 0&0&1&1&1&0&0&1&1&0&1&1&1&0&0&1&1&0&1&1&0&0&0&0&0 \end{bmatrix}.$$

Next, we use the fourth spectrum (subscript 3) as a query. After performing the same preprocessing, the query is divided into 25 segments. We calculate the distance between each segment of the query and all the cluster centroids in the current segment, and we get a distance table of 25*2 in which the rows represent the segments (0–24) and the columns represent the

$$\text{cluster centers (0, 1): } \begin{bmatrix} 1.111 & 0.237 \\ 0.258 & 1.259 \\ \vdots & \vdots \\ 0.198 & 0.742 \end{bmatrix}.$$

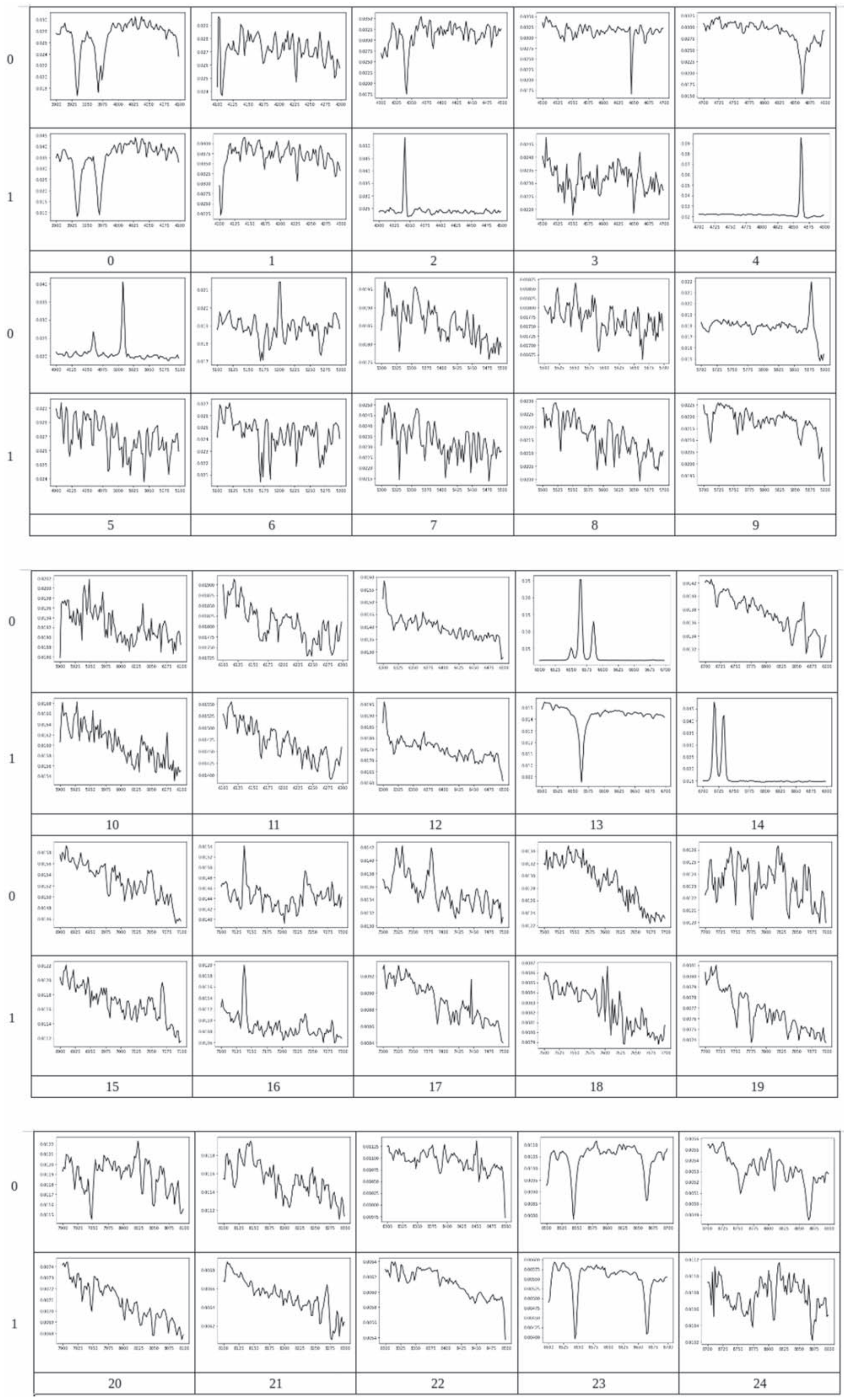

Figure 2. We get 50 cluster centers after we divide the four spectra shown in Figure 1 into 25 segments and cluster the segments into two clusters using the K-means method. The numbers arranged horizontally are the subscripts of the segments (0–24), and the vertical numbers on the left are the flags of the cluster centers (0, 1).

According to the cluster to which each segment of each spectrum belongs, we can quickly get the sub-distance of the query spectrum on this segment from the current spectrum (this distance calculation method is an asymmetric distance). Then we connect the 25 sub-distances through the Cartesian product to get the approximate distance between the two. The spectrum with the smallest approximate distance to the query is the nearest neighbor spectrum of the query. For example, if the compressed data for one spectrum is $0\ 1\ \cdots 0]$, the distance between the query and this spectrum can be obtained by $1.111 + 1.259 + \cdots + 0.198$.

After calculation, we get the distances between the fourth spectrum and all the four examples (with subscripts 0, 1, 2, 3) which are: 4.423, 15.430, 8.656 and 3.241. We can see that the nearest neighbor for the query is the spectrum with subscript 3 because of the minimum distance. The nearest neighbor found is the query itself, that is for sure. The distance between the query and the nearest neighbor is not 0. This is because our distance calculation is based on the cluster centroids, not on the direct distance between the two. We use Euclidean distance throughout the work.

3.1.4. Validity of PQ Method in Spectral Retrieval

We select 100 galaxy spectra with strong emission lines released in LAMOST DR6 and 4000 F-type spectra with S/N in *g*-band greater than 50 in LAMOST DR5, and use the PQ retrieval method to search for the nearest neighbors of the galaxy spectra. By setting different segments and the number of cluster centers, the PQ method is used to test the effectiveness of spectral data retrieval. This experiment uses the same wavelength range, step size setting, and normalization process as Section 3.1.3. In the experiment, we set the segment to 25, and the number of cluster centers to 2, 3, and 4 respectively. The results show that the proportions of galaxy spectra searched out are 58%, 87%, and 96% respectively. So we can see that the PQ method is effective in the retrieval of spectral data.

3.2. Work Strategy

In our main work, we use the spectra of "GALAXY" released in LAMOST DR6 as the templates, and the spectra of "Unknown" in LAMOST DR7 as the unresolved spectra. We perform 12 approximate nearest neighbor retrievals for picking out spectra of galaxies with strong emission lines in the unsolved spectra. The retrieval strategy generates queries for every item of the unsolved spectra and uses the templates as the objects to be retrieved. The retrieval process is shown in Figure 3.

When executing the strategy, all published spectra of "GALAXY" are used as templates which are noted as T1 in Figure 3, and this retrieval work occurring in T1 is represented by the flow direction of the real arrow in Figure 3. We use the emission line characteristics of spectra of galaxies as selection criteria to pick out the candidate unresolved spectra with strong

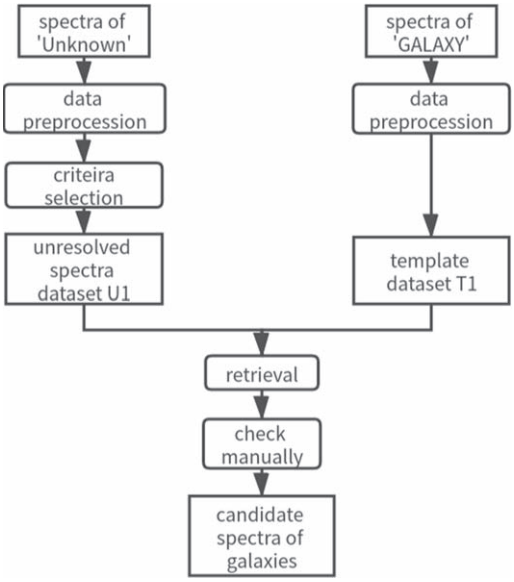

Figure 3. Retrieval process for spectra of galaxies having emission lines. and the program flow is represented by solid arrows.

emission lines in the "Unknown" spectra. We note the set of candidate spectra as U1 in Figure 3. And we calculate the corresponding redshift values using CL and RT methods (Section 2.2.4). Then we generate a query for each spectrum of these candidates in U1, and use the PQ based approximate nearest neighbor method to retrieve all template spectra in T1 to find the nearest neighbor of each candidate spectrum. We check the error between the calculated redshift of the candidate spectrum and the published redshift of the nearest neighbor template spectrum. If the error is within the threshold, a potential target spectrum is obtained along with its redshift. All of these targets obtained using this threshold make up our result set as shown in Figure 3. Finally, we visually inspect the result set to further determine the candidate galaxy spectra.

4. Method Application in Spectra Data set of LAMOST

4.1. Main Process of Our Work

Our main process of retrieving spectra of galaxies with strong emission lines from the "Unknown" spectral data set of LAMOST can be displayed as the following Process I:

1. Retrieval Process
   (a) First, we pick out spectra with strong emission lines using the SW and CL methods in 492,938 "Unknown" spectra of LAMOST DR7 after performing interpolation and normalization on them, and get 123,693 spectra with computed redshifts.
   (b) Second, we use the PQ method to retrieve the approximate nearest neighbors of each of the 123,693 spectra. We perform all queries on the

Table 1
12 Combinations of three Parameters in 12 Operations in our Retrieval Work

step (Å)	Segment	Cluster_K
2	25	100
2	50	200
2	100	400
2	250	800
2	500	1600
2	1250	3200
1	50	200
1	100	400
1	200	800
1	500	800
1	500	1600
1	1000	3200

177,270 "Galaxy" spectra of LAMOST DR6 used as the template set. The process is detailed as follows:

i. We perform the same interpolation and normalization on the template spectra, then segment the data points in a fixed number and perform a certain number of clusters on segments. The PQ model keeps the data information of each cluster center. We define the codeword according to the cluster centers of each segment of each template spectrum, and connect all codewords through the Cartesian product to form a codebook of all the template spectra for each run.

ii. For each one of the 123,693 spectra, we can get a table composed of the distance from it to each cluster center. When we query in the codebook, we can quickly get the distances between this current spectrum and all the templates. We only choose the minimum distance, and in each operation we get the nearest neighbor template of each spectrum.

iii. We operate the PQ retrieval 12 times and get 12 independent result files.

(c) Third, from the 12 files, we select the spectra of unique sources whose computed redshift values are in good agreement with the released ones of the corresponding templates. We merge the results and get 16,188 candidate galaxy spectra.

(d) Finally, after visual inspection, we keep 10,266 spectra with obvious emission lines as our target galaxy candidates.

2. Detailed implementation

We explain the detailed implementation of our retrieval process using Table 1. There are three parameters used in the retrieval process: step, Segment, Cluster_K. The first column shows the interpolation step size we set on each spectrum. Step 2 Å represents the operation with 2500 interpolation points (the entire wavelength range is 3900–8900 Å) and 1 Å represents the operation of 5000 points (Section 2.2.1). The second column shows the number of segment we cut. The Cluster_*K* value shows the number of cluster centers for each run of clustering. For example, we set step to 2 Å, Segment to 25, Cluster_K to 100. That is, we segment every normalized spectrum having 2500 interpolated points to 25 parts, and the same K-means method (the number of cluster centers is 100) is operated on these 25 segments respectively. Thus we can build up the PQ model and get the distance table which can be used to retrieve the nearest neighbor. We pick out only those with strong emission lines, and only retain the spectra whose calculated redshift values agree with the released redshift values of the template spectra (nearest neighbors). In this way, we can get the candidates under the combination of these three parameters (step:2 Å, Segment:25 and Cluster_K:100) from the "Unknown" data set. We run the similar processions to get their respective results for other combinations in Table 1, and count the cumulative sum value of the results.

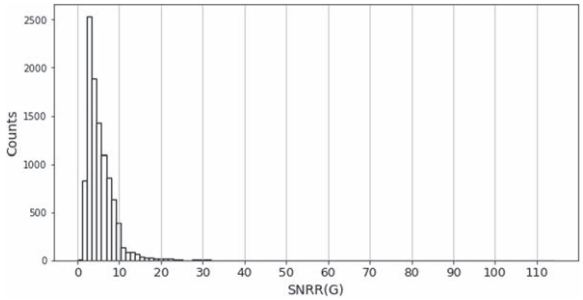

Figure 4. The distribution of S/N in r-band for our candidate spectra in LAMOST "Unknown" data set, and obviously most of our spectra having S/N lower than 10, which shows our method can deal with low-quality spectra correctly.

4.2. Results Analysis of Our Work

From the above operation we totally get 10,266 candidate spectra with obvious emission line in LAMOST DR7 "Unknown" spectral data set. In this section, we will analyze the properties of our result spectra.

4.2.1. Properties of Candidate Spectra

Figure 4 shows the distribution of S/N in r-band for our candidate spectra ("G" in the label means the result of galaxy spectra), and we can find most of them are low-quality spectra with S/N lower than 10. This may be one reason why some spectra are mis-classified to "Unknown" data set. It can also prove that our method can deal with low-quality spectra correctly and it also has a good applicability. We show the magnitude distribution in *g*-band for our candidate spectra in the Figure 5 ("G" is same to Figure 4), in which there are both bright and dark sources.

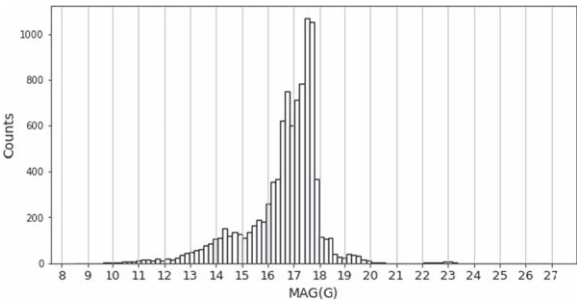

Figure 5. The magnitude distribution of our candidate spectra shows that there are both bright and dark sources.

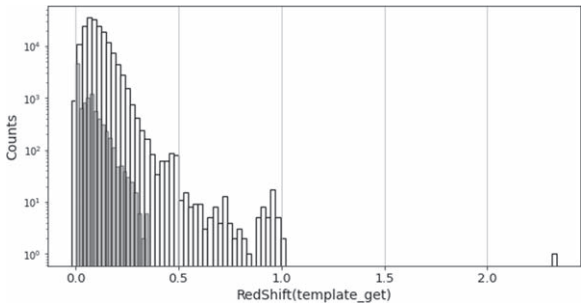

Figure 6. The gray rectangles in the foreground is the distribution of redshift values for candidate spectra in our result, while the white rectangles in the background display the distribution of redshift for the whole template data set.

The distribution of redshift values (z: $0 \sim 0.35$) for our candidate spectra in LAMOST DR7 "Unknown" data set is shown as gray rectangles in the foreground in Figure 6, and the white rectangles in the background display the distribution of redshift for our template spectra of "Galaxy". This shows that the galaxy spectra that we get are low redshift.

4.2.2. Spectra of Ionized Nebula in the Galaxy

According to the redshift distribution in Figure 6, we find that some of the candidate spectra have small redshift values. After examining these spectra, we find some spectra show strong characteristics of nebula emission-line. We try to look for Galactic H II regions with these spectra. First, with these 10,266 candidate spectra, we select spectra whose calculated redshift and corresponding template published redshift are both less than 500 km s^{-1} divided by the speed of light, and get 4438 spectra.

After cross matching with HIICat_V2, we get 985 spectra which fall in a total of 72 H II regions given by WISE. Of these H II regions, 43 (901 spectra) have been identified and marked as "K" which means 'known H II regions', and the other 29 (84 spectra) are newly certified H II regions in our work.

Then we try to spectroscopically confirm these 4438 spectra with an emission-line diagnostic criterion based on $\log([\mathrm{S\,II}]\lambda6718, 6732/\mathrm{H}\alpha)$ and $\log([\mathrm{N\,II}]\lambda6585/\mathrm{H}\alpha)$, using the following equation: $\log(\frac{[\mathrm{S\,II}]}{\mathrm{H}\alpha}) \geqslant 0.63\log(\frac{[\mathrm{N\,II}]}{\mathrm{H}\alpha}) - 0.55$ (Riesgo-Tirado & López 2002; Magrini et al. 2003; Kniazev et al. 2008). Through calculation, we get 1082 spectra which meet the diagnostic criteria of H II regions. Of these spectra, 285 spectra fall in H II regions given by WISE, and 12 H II regions (16 spectra) are identified spectroscopically that marked "G", "C" or "Q" in WISE. We newly find 797 (1082–285 = 797) spectra that may be in the H II regions of the Galaxy.

We provide a catalog in electronic form in the online version of these H II region spectra (985+797 = 1782), as shown in Table 2, containing the filename of LAMOST, WISE name (If the position represented by the spectrum falls in the H II regions of WISE catalog, here is the name of WISE, otherwise here is only the "-"). The number marked by brackets after the name indicates the potential 29 candidate H II regions of the Galaxy that we have identified in our work. If the name is finally marked with an asterisk, it indicates the 12 (out of 29) regions of the Galaxy H II region obtained using the diagnostic formula. The other information in the catalog is catalog type in WISE, R.A. and Decl. given by WISE, $\log(\frac{[\mathrm{N\,II}]}{\mathrm{H}\alpha})$, $\log(\frac{[\mathrm{S\,II}]}{\mathrm{H}\alpha})$, and New_found (1 represents spectrum newly found that may be located in H II regions in the Galaxy, 0 represents spectrum that is located in H II regions of WISE). The distribution of velocities of our 1782 H II region spectra is shown in Figure 7 and we can see that most velocities are less than 150 km s^{-1}. Among 4438 spectra with a velocity of less than 500 km s^{-1}, there are still 2656 spectra that cannot be clearly classified, though they have obvious emission lines and small redshift.

4.2.3. Classification of Spectra of Galaxy Candidates

There are 5828 (10,266–4438) spectra having velocities larger than 500 km s^{-1} divided by the speed of light. We classify them to the four classes: Star-forming, composite, LINER (Heckman 1980) and Seyfert (Baldwin et al. 1981), using the criteria that the emission line features of galaxy spectra contain significant information bearing on the stellar populations (Morgan & Mayall 1957; Wang et al. 2018b). When classifying, we use the line strength ratios $\log([\mathrm{O\,III}]\lambda5008/\mathrm{H}\beta)$ and $\log([\mathrm{N\,II}]\lambda6585/\mathrm{H}\alpha)$ (Baldwin et al. 1981), and the classification is shown in Figure 8. There are 1624 spectra of Star-forming galaxies, 1945 composite galaxies, 1725 LINERs and 534 Seyferts in our work. We provide a catalog in electronic form of classification of the 5828 galaxy spectral candidates, as shown in Table 3 (Column 6).

By cross-matching our 5828 spectra with Strasbourg astronomical Data Center (CDS) catalog and Sloan Digital Sky Survey (SDSS), we get 4245 spectra from CDS and 4254 spectra from SDSS. The total number of same source spectra from CDS and SDSS is 4682, corresponding to 4573 unique sources. Our whole 5828 spectra of galaxy correspond to 5701 unique sources. Then for the other unique sources we make search in the NASA/IPAC Extragalactic Database (NED). We get 1022 unique sources which matching with 1038 spectra of

Table 2
Catalogue of H II Region in our Result

file_LAMOST	WISE_Name	Catalog	GLong(deg)	GLat(deg)	Radius(arcsec)	log(N II_H)	log(S II_H)	New_found
spec-56649-GAC078N31M1_sp12-177.fits	G173.588-01.606	K	173.588	−1.606	1298	⋯	⋯	0
spec-57779-GAC099N04B1_sp14-053.fits	G206.316-02.103	K	206.316	−2.102	3133	⋯	⋯	0
spec-57779-GAC099N04M1_sp03-066.fits	G206.316-02.103	K	206.316	−2.102	3133	⋯	⋯	0
spec-57044-GAC057N34M1_sp11-028.fits	G159.957-12.738	K	159.957	−12.737	5689	⋯	⋯	0
spec-57071-GAC084N35M1_sp09-179.fits	G173.468+03.230[1]*	G	173.468	3.23	1369	−0.444	−0.45	0
spec-56649-GAC078N31M1_sp15-048.fits	G173.156-03.442 [2]	Q	173.156	−3.442	728	⋯	⋯	0

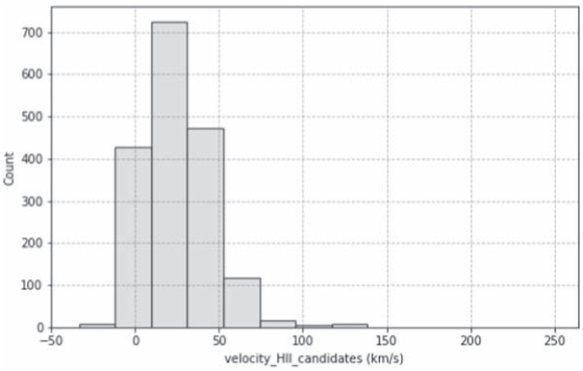

Figure 7. The distribution of velocities shows that most velocities of the H II region spectra in the Galaxy of our work are less than 150 km s^{-1}.

LAMOST. There are 108 spectra (corresponding to 106 unique coordinates of R.A. and decl.) that have no same source in the three databases. Thus we process all the 5828 spectra (4682 +1038+108 = 5828) and the results exactly match with the 5701 unique sources (4573+1022+106 = 5701). Table 3 shows our results. The columns in the table respectively represent information related to LAMOST (Columns 1–6), SIMBAD information (Columns 7–9), SDSS information (Columns 10–14), and contents related to NED (Columns 15–18). The first group of information includes file name of LAMOST, R.A., Decl. flag of newly discovery (1 for new, 0 for existing), the calculated redshift of our work, and the galaxy classification obtained using the line-strength ratio. Information related to SIMBAD includes identity document, object classification and redshift. The third group of information includes identity document of SDSS, object classification, photometric light source radius, redshift and redshift errors. Information related to NED includes name, category, redshift and redshift error.

In published general catalog of LAMOST, there is a keyword named "OBJTYPE", which shows the type of source selection target given by different researchers in the process of source selection. Figure 9 shows the distribution of "OBJTYPE" for our spectra of galaxy candidates. We can find that a large proportion of the selected targets are star like targets, which may be one reason for mis-classifying galaxy spectra into the "Unknown" spectra.

We get 5720 (4682+1038 = 5720) spectra after cross-matching our galaxy candidates with the three databases CDS, SDSS and NED, and these spectra represent 5595 (4573 +1022 = 5595) different galaxies. Redshift is an important attribute of galaxies and can help study the motion properties of galaxies and the expansion of the universe. Comparison of our calculated redshifts with the ones in the three databases CDS, SDSS and NED is represented in Figure 10. We can see that our result agree with the literatures, and this can prove the applicability and effectiveness of our method. However, some sources in the figure have large redshift errors, which should be one of the research contents of our follow-up work.

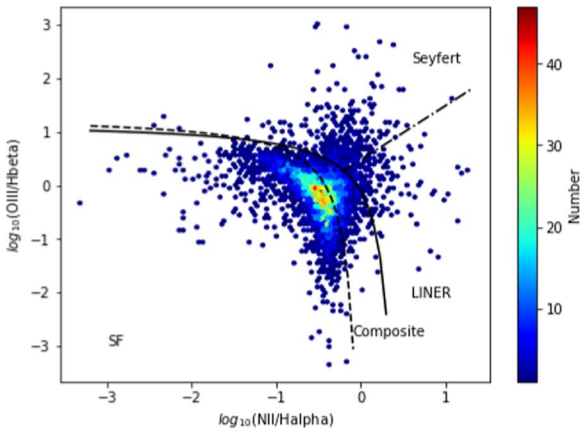

Figure 8. The classification for the galaxy candidate spectra having velocities larger than 500 km s^{-1}. We get 3503 candidate spectra of Star-forming galaxies, 1254 spectra of composite galaxies, 191 of LINERs, and 880 of Seyferts.

We show the distribution of celestial type for same sources of our galaxy targets and CDS, SDSS and NED in Figure 11. From the distribution map, we can find that the types of our galaxies in the literature are diverse. Some of the interesting celestial bodies can also be part of our follow-up research.

5. Discussion and Conclusion

The LAMOST 1D pipeline uses the chi-square minimum method to match the target spectrum with the template. However, the type of template spectrum in the pipeline is limited, such as the lack of extragalactic galaxy spectrum. In addition, some galaxy spectra are difficult to classify because of the low S/N. They are released as "Unknown" spectra, although some of these spectra have obvious emission lines. To quickly find out them from the several hundred thousand spectra, we employ a retrieval method based on the product quantization method in this paper. Different from the traditional template matching approach, we use galaxy spectra released by LAMOST as templates and proceed with a large-scale nearest neighbor retrieval. Using the essence of segment quantization coding of the PQ method, we can complete this process in a short time.

In this work, we select 500 km s^{-1} as the velocity threshold to distinguish extragalactic sources from sources of the Milky Way, which is consistent with the classification standard of the LAMOST 1D pipeline. Spectra with a radial velocity of greater than 500 km s^{-1} are extragalactic source candidates. The classification results of these extragalactic sources are given in Table 3. We cross-match sources with a velocity of less than 500 km s^{-1} with WISE H II region catalog (Anderson et al. 2015, 2018), and spectroscopically confirm spectra with emission-line diagnostic criteria. We get 1782 spectra falling in the Galactic H II regions in our work, as shown in Table 2. Among the spectra with a velocity less than 500 km s^{-1} in our results, there are still 2656 spectra whose categories have not

Table 3
Catalogue of Candidates of Galaxy Spectra Found in our Result

file_LA	ra	dec	newG	z_our	class_our	simbad_id	simbad_type	simbad_z	SDSS_id	SDSS_type	SDSS_prtroRad	SDSS_z	SDSS_zErr	NED_name	NED_type	NED_z	NED_zErr
spec-57721-M31025N38M1_sp11-034.fits	26.314	40.798	0	0.121	SF	_	_	_	_	_	_	_	_	WISEA J014515.26+404753.3	IrS	_	_
spec-55886-F8606_sp14-059.fits	52.736	6.035	0	0.100	SF	_	_	_	_	_	_	_	_	WISEA J033056.58+060206.7	IrS	_	_
spec-57389-M31019N33M1_sp13-112.fits	21.210	33.917	0	0.231	SF	_	_	_	_	_	_	_	_	WISEA J012450.62+335502.6	UvES	_	_
spec-57055-HD133001N190343B01_sp03-109.fits	201.099	19.538	0	0.071	SF	2MASX J13242373+1932179	Galaxy	0.071	1237668272434774166	GALAXY	7.914	0.071	1.3E-05	_	_	_	_
spec-57491-HD124702N090350M01_sp16-118.fits	190.947	10.800	0	0.106	SF	LEDA 1384624	Galaxy	0.106	1237658493894656144	GALAXY	8.388	0.106	1.6E-05	_	_	_	_
spec-57397-HD130439N535127M01_sp15-167.fits	195.205	55.171	0	0.086	SF	LEDA2492960	Galaxy	0.086	1237658802577408090	GALAXY	7.618	0.086	1.5E-05	_	_	_	_

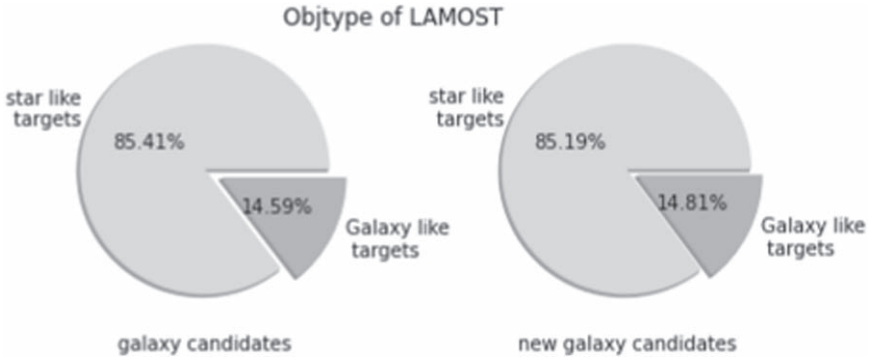

Figure 9. The proportion of values of "OBJTYPE" for our galaxy spectra, which shows the type of source selection target. We can see a large proportion of the selected targets is star like targets, which may be the reason for misclassification.

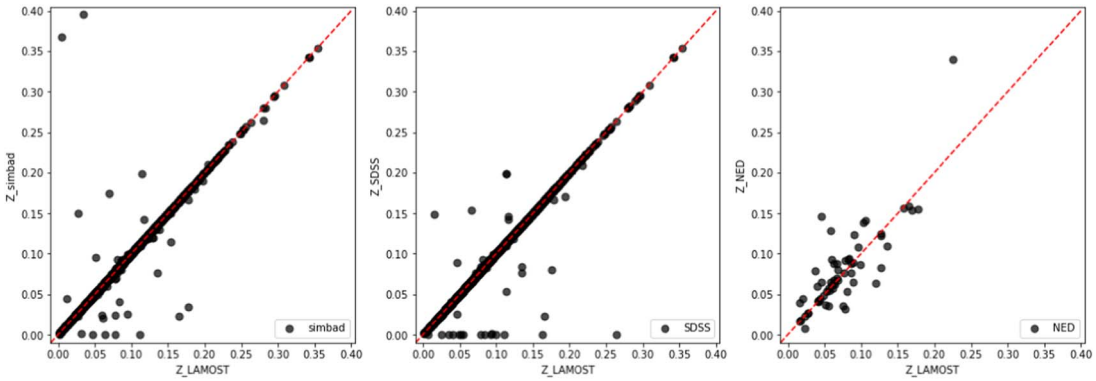

Figure 10. Comparison of our calculated redshifts with the ones in the three databases CDS, SDSS, and NED. The red dashed lines represent 1:1 lines.

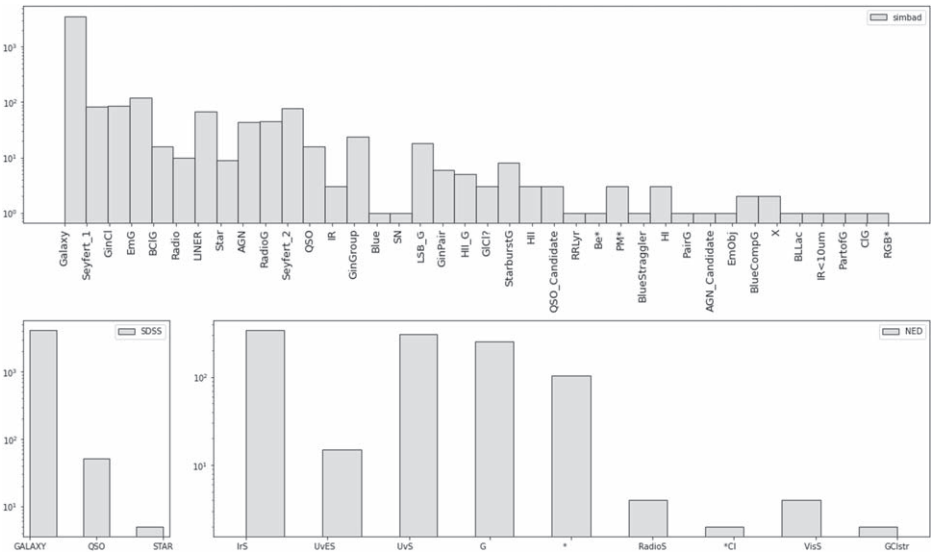

Figure 11. The distribution of values for keyword "main_type" for same sources between our galaxy targets and CDS catalog, keyword "class" for SDSS and keyword "ObjectType" for NED.

been determined, and their corresponding sources need more observation results to support detailed research.

To conclude, using released spectra of LAMOST as templates, and using PQ-based approximate nearest neighbor retrieval as the tool, we have retrieved 10,266 potential galaxy spectra in the LAMOST "Unknown" spectral data set. Among them, we get 985 spectra which fall in 72 confirmed H II regions proposed by WISE. We newly find 797 spectra in the H II region of the Galaxy by calculation. Most of these 1782 ($985+797=1782$) spectra have radial velocities less than 150 km s^{-1}. We also get 5828 spectra of galaxies with radial velocities larger than 500 km s^{-1}, and we divide them into four categories. By crossing match the 5828 spectra with CDS, SDSS, and NED, we find 4245 existing in CDS, 4254 in SDSS, 1038 in NED. We get 108 spectra (corresponding to 106 unique coordinates of R.A. and decl.) that are not collected in the three databases. All these objects will be archived as galaxies and H II regions in the following release version, LAMOST DR8. The method used in this paper will be used as a subsequent supplement to the LAMOST 1D pipeline.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) under Nos.U1931209, 11903008, U1931106, and the Science Foundation of DeZhou University (grant No.2019xjrc39). Guo Shou Jing Telescope (the Large Sky Area Multi-Object Fiber Spectroscopic Telescope, LAMOST) is a National Major Scientific Project built by the Chinese Academy of Sciences. Funding for the project has been provided by the National Development and Reform Commission. LAMOST is operated and managed by the National Astronomical Observatories, Chinese Academy of Sciences.

References

Anderson, L. D., Armentrout, W. P., Johnstone, B. M., et al. 2015, ApJS, 221, 26
Anderson, L. D., Armentrout, W. P., Luisi, M., et al. 2018, ApJS, 234, 33
Anderson, L. D., Bania, T. M., Balser, D. S., et al. 2014, ApJS, 212, 1
Anderson, L. D., Bania, T. M., Balser, D. S., & Rood, R. T. 2011, ApJS, 194, 32
Baldwin, J. A., Phillips, M. M., & Terlevich, R. 1981, PASP, 93, 5
Blanton, M. R., Bershady, M. A., Abolfathi, B., et al. 2017, AJ, 154, 28
Brinchmann, J., Charlot, S., White, S. D. M., et al. 2004, MNRAS, 351, 1151
Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, RAA, 12, 1197
Deng, L.-C., Newberg, H. J., Liu, C., et al. 2012, RAA, 12, 735
Eisenstein, D. J., Weinberg, D. H., Agol, E., et al. 2011, AJ, 142, 72
Esteban, C., Fang, X., García-Rojas, J., & Toribio San Cipriano, L. 2017, MNRAS, 471, 987
Helou, G., Madore, B. F., Schmitz, M., et al. 1995, The NASA/IPAC Extragalactic Database (Dordrecht: Springer Dordrecht)
Heckman, T. M. 1980, A&A, 87, 142
Jégou, H., Douze, M., & Schmid, C. 2011, ITPAM, 33, 117
Kauffmann, G., Heckman, T. M., White, S. D. M., et al. 2003, MNRAS, 341, 33
Kennicutt, R. C. J. 1992, ApJS, 79, 255
Kniazev, A. Y., Pustilnik, S. A., & Zucker, D. B. 2008, MNRAS, 384, 1045
Luo, A. L., Zhang, H.-T., Zhao, Y.-H., et al. 2012, RAA, 12, 1243
Luo, A.-L., Zhao, Y.-H., Zhao, G., et al. 2015, RAA, 15, 1095
Magrini, L., Perinotto, M., Corradi, R. L. M., & Mampaso, A. 2003, A&A, 400, 511
Morgan, W. W., & Mayall, N. U. 1957, PASP, 69, 291
Napolitano, N. R., D'Ago, G., Tortora, C., et al. 2020, MNRAS, 498, 5704
Ochsenbein, F., Bauer, P., & Marcout, J. 2000, A&AS, 143, 23
Riesgo-Tirado, H., & López, J. A. 2002, RMxAA, 12, 174
Veilleux, S., & Osterbrock, D. E. 1987, ApJS, 63, 295
Wang, L.-L., Luo, A. L., Hou, W., et al. 2018a, PASP, 130, 114301
Wang, L.-L., Luo, A. L., Shen, S.-Y., et al. 2018b, MNRAS, 474, 1873
Xiang, M. S., Liu, X. W., Yuan, H. B., et al. 2015, MNRAS, 448, 822
Yip, C. W., Connolly, A. J., Szalay, A. S., et al. 2004, AJ, 128, 585
Zhao, G., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, RAA, 12, 723