Journal of Science and Technology
ISSN: 2362-1443 (Print) 3059-9490 (Online) 2026, Vol. 5, No. 1

Information Retrieval from Job Posts based on K-means++ Clustering Algorithm

Sinuna Chaudhary
Nepal College of Information Technology,
Pokhara University, Nepal
sinuna.chaudhary@gmail.com

Shashidhar Ram Joshi*
Pulchowk Campus, Institue of Engineering
Tribhuvan University, Nepal
srjoshi@ioe.edu.np

*Corresponding author

***Abstract*—The research paper deals with two main sections: firstly, the experiment comparison between k-means and k-means++ have been done using Elbow method and Silhouette method. Since, K-means++ is better than K-means, this research tries to justify that K-means++ has higher performance than K-means. Secondly, K-means++ has been used for Search and Information Retrieval system. Information Retrieval is an activity to obtain information system resources that are relevant to an information need from a collection of those resource. This research is useful to retrieve relevant documents that match a given query. When user add input such as industry type, job types, skills, and state, it will automatically calculate average and display the ranking. Subjective evaluation with DCG(Discounted Cumulative Gain) is done in order to measure ranking quality of information retrieval.**

***Keywords* — Information Retrieval, Machine Learning, K-means, K-means++, Elbow Method, Silhouette Analysis, Discounted Cumulative Gain Information Retrieval, Machine Learning, K-means, K-means++, Elbow Method, Silhouette Analysis, Discounted Cumulative Gain:**

I. Introduction

With the growth of online jobs, the job demand is increasing rapidly. The job demand is increased due to the availability of internet technologies. Providing relevant job information to user according to the user's query is challenging task. Web is one of the important source of a huge collection of information resources and mining of knowledge from huge amount of data is complex task. This research is useful in search engines and the information retrieval field and it is especially useful for recruitment website.

Information retrieval is the field of computer science that deals with retrieving relevant documents that match a given query. It is the process of finding relevant information from a collection of information resources as fast as possible.

In this research, clustering algorithm is used in information retrieval for document grouping. Clustering is an unsupervised Machine Learning task that automatically divides the data sets into clusters, or group of similar items. Cluster analysis is widely used for exploratory data analysis to find hidden pattern or grouping in data. The objective of the unsupervised Machine Learning technique is to find similarities and group similar data points together [1]. The algorithm calculates the distance between each point from the centroid of the cluster. Cluster analysis is also widely used for exploratory data analysis to find hidden pattern or grouping in data. The advantage of clustering over classification is that it helps features to distinguish different groups.

Unsupervised learning algorithm has emerged as the most effective technique for discovering patterns in data. The main reason to use unsupervised learning is because it is useful technique for automated information retrieval from job posts wherein the unstructured jobs data can be segregated into clusters and used to identify certain patterns which leads to a more customized approach. K-means++ clustering algorithm is used to cluster the data sets in this research work. It is the versatile and smart enough to correctly cluster the unstructured data. This technique comprises machine learning algorithm through which data analysis can be drawn inferences from data sets without labelled responses. Data points are clustered based on feature similarity.

Finding relevant job from the pool of job posts is a challenging task. Manual scanning and sorting through jobs is tedious. The Machine Learning technique can be implemented to find the most relevant job from the job posts automatically. Data sets of 5483 is applied for this research. Firstly, K-means++ algorithm is implemented for this system work. Secondly, information retrieval method is implemented and the relevant documents are retrieved according to the given query. It helps to find the most relevant job from a collection of job posts.

In order to get the right solution, a clearly defined objectives are necessary. The main objectives of the research work are listed below:

- To cluster the job similarities.
- To Find out the most relevant jobs from the pool of job post

This research studies the information retrieval from jobs based on K-means++ clustering algorithm. The main goal of this research is to find the most relevant job for job seeker from

DOI : https://doi.org/10.3126/jost.v5i1.93045
Article History: Received: 15 Nov. 2025, Revised: 14 Jan. 2026, Accepted: 26 Feb. 2026.

pool of job posts. The scope of this research are discussed below:

- Applicant pools have been growing as a result of online job posting, which make it easier for applicant to find jobs online and apply to them with ease.
- This research is useful in the information retrieval field and it is especially useful for recruitment website.

II. Related Works

This research [2] presents Using k-means++ algorithm for researchers clustering. The clustering of researchers based on publication is one of identifying community of researchers. This helps researchers know the relationship with other researchers regarding similarities of topics and disciplines of publications based on the research community.

This research [3] presents Movie ranking using k-means clustering algorithm in data mining. K-means clustering algorithm is used to find the ranking from given user information available on social network website like orkut, facebook, and twitter. K-means helps to reduce work complexity. When user enters new comments about movie, it will automatically calculate average and display the ranking.

This research [4] presents Comparative study of k-means and k-means++ clustering algorithms on crime domain and compares two approaches in crime document clustering. The experimental results on crime datasets identifies the best seed for initial cluster centers. Comparison on k-means and k-means++ have been done and it shows that k-means++ works significantly better than k-means.

This research [5] presents A better k-means++ Algorithm via Local Search. This research obtain the result by a simple combination of k-means++ sampling with a local search strategy. The algorithm evaluation is done empirically and shows the quality of solution in practice. This research proposed a simple variation of k-means++ algorithm based on local search and prove that the algorithm achieves a constant factor approximation. The experiment shows the effectiveness of this method.

The research [6] presents Personalized Travel Recommendation Using K-Means Clustering. It provides personalized travel sequence recommendation with the help of travelogue and community contributed photos. Travelogue websites have good content about landmarks and experience which are written by users. These data are useful for reliable POIs (Point of interest) mining, travel routes mining. It recommend personalized travel POIs and routes based on user's interest.

III. Research Methodology

This research works on Machine Learning algorithm : K-means++ clustering algorithm. It is the popular cluster analysis method, aim is to partition N data points into k clusters with the nearest mean.

Formula:

$$J(V) = \sum_{i=1}^{c} \sum_{i=1}^{ci} (\|xi - vj\|)^2 \quad (1)$$

where,

$$||xi - vj||$$

is the Euclidean distance between xi and vj.
ci is the number of data points in ith cluster.
c is the number of cluster centers.[7]

Here, the clustering technique has partitioned the entire data points into twenty clusters. The data points within a cluster are similar to each other but different from other clusters.

A. Conceptual Model

The experimental framework build a significant reference for unsupervised learning, more precisely made of 20 clusters (0 to 19) gathering 5483 data sets. Finally, evaluation of clustering and analyze the results of application.

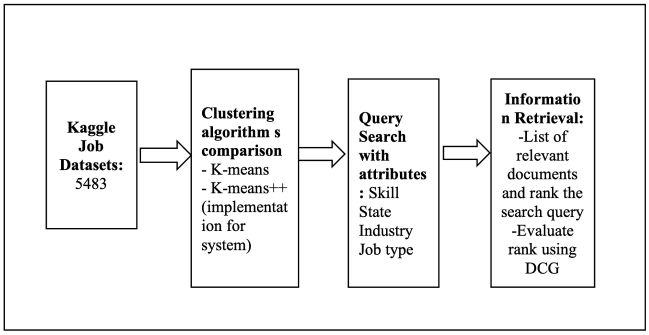

Fig. 1. Conceptual Model

The working mechanism for this system is explained below:

- The data sets is downloaded from Kaggle Website. The downloaded url is: https://www.kaggle.com/elroyggj/indeed-dataset-data-scientistanalystengineer
- K-means and K-means++ clustering algorithms are compared for justification as K-means++ has higher performance than K-means. So, K-means++ is used for the system implementation.
- The process of information retrieval starts when a user creates query with skill sets, industry type, job type, and location and these user defined queries are the statements of needed information.
- The information retrieval system generates several collections of data objects from which the most relevant documents are taken into consideration. The ranking of relevant documents is done to find out the most relates document to the given query.

B. Flow Chart

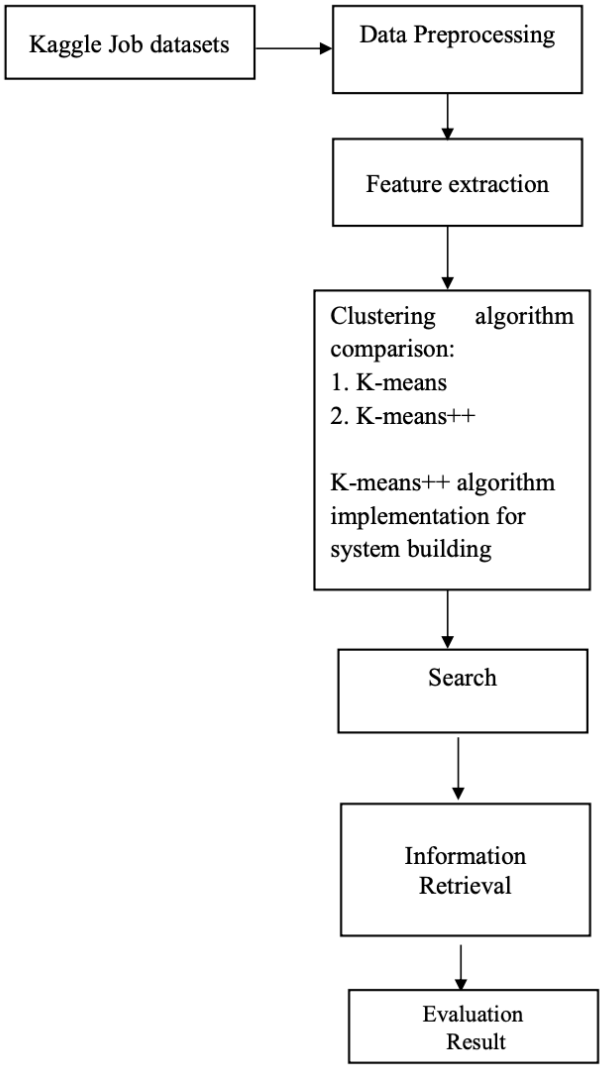

Fig. 2. Flow Chart

*C. **Data Collection***

For the experiment purpose, the data sets is gathered and downloaded from Kaggle website (https://www.kaggle.com/elroyggj/indeed-dataset-data-scientistanalystengineer) [8]. The data sets is of indeed website, an American search engine for job listing which contains 5483 rows with 15 attributes including: Job title, Link, Salary, Job type, Skill, No. of skill, Company, No. of reviews, No. of stars, Date since posted, Description, Location, Company revenue, Company employee, and Company industry.

The following are graphs of the particular data sets. Job type, Skill, Industry, and Location are the important column field for this research work.

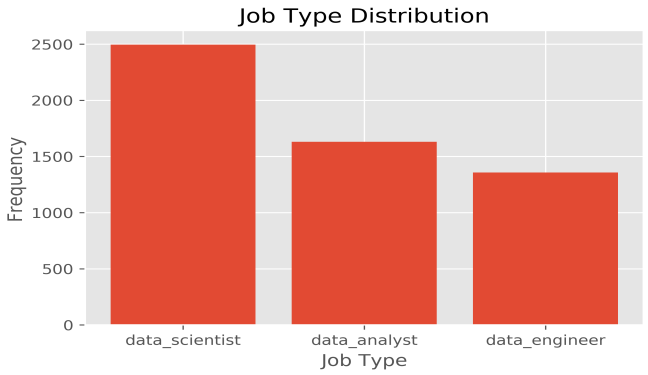

Fig. 3. Job type

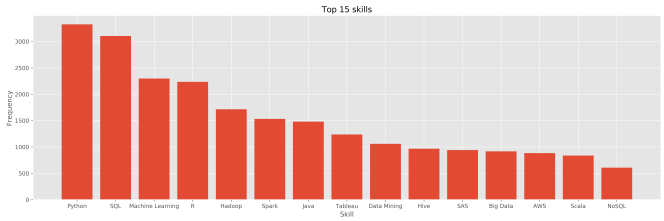

Fig. 4. Skills

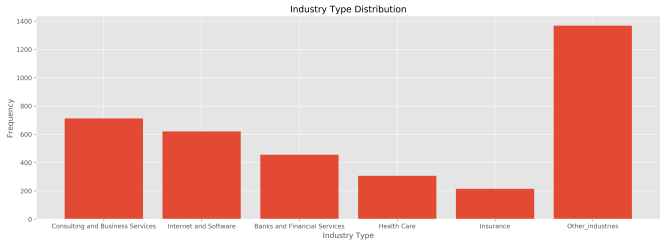

Fig. 5. Industry

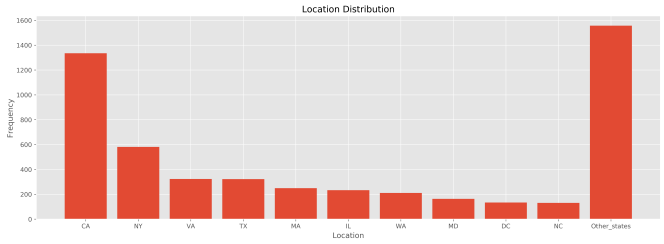

Fig. 6. Location

*D. **Data Prepossessing***

Data cleaning is one of the important step in data pre-processing. Data cleaning is done with python library - pandas for detecting missing values. Data prepossessing is the technique that involves transforming raw data into an understandable format. Real world data is often incomplete and inconsistent and/or lacking in certain behaviors or trends, and is likely to contain behaviors or trends, and is likely to contain many errors. The data preprocessing consists of the

following:

- Droping unnecessary columns in a Data Frame
- data.dropna for skipping unnecessary rows in a CSV file
- Total datasets

  `5715->5483`

  (after dropping the datasets)

This data pre-processing includes following:

- Data Cleaning
- Normalization
- Transformation
- Feature selection and extraction

E. Feature Extraction

Feature extraction is one of the important step in pre-processing.

The following features were extracted: -

1. Job type
2. No. of star
3. No. of skill
4. No. of review
5. Industry type

Job type and Industry type are extracted using One Hot Encoding whereas no. of star, no. of skill, and no. of review are extracted using Normalization.

1) One Hot Encoding:

One Hot Encoding is the most widespread standard approach for categorical data. It creates new (binary) columns, indicating the presence of each possible value from the original data [12]. The values in the original data are Data scientist, Data analyst, Data engineer. It create a separate column for each possible value. For example, wherever the original value was Data scientist, we put a 1 in the data scientist column.

2) Normalization (Min/Max)

Min Max is a technique that helps to normalize the data. It consists of scale the data between 0 and 1 [13]. Normalization helps understand the data easily. No. of star, no. of skill, and no. of review features are extracted using Min-Max Normalization. [11]

It is defined as :

$$MinMax = \frac{V - min}{max - min}(newMax - newMin) + newMin \quad (3)$$

Where,
Min: min value of given attribute.
Max: max value of given attributes
newMax = 1 and newMin =0

F. Comparison: K-means and K-means++ Clustering

K-means++ algorithm is the variation of the standard k-means algorithm. The main reason to use these algorithm is because they are useful technique for information retrieval wherein the unstructured jobs data can be segregated into clusters and used to identify certain patterns which leads to a more customized approach. K-means starts with allocating cluster centers randomly and then looks for better solution whereas, K-means++ is useful for improvement of the centroid initialization for k-means. So, both algorithms use random initialization as a starting point, so can give different results on different runs [9]. The experiment analysis is done comparing K-means and K-means++ algorithm. The comparison is done for justification to use K-means++ algorithm. This justifies why K-means++ is better than K-means.

To run a k-means algorithm, the process started with initialize random data centroid K= 20. So, that the data were grouped into 20 clusters. This technique comprises machine learning algorithms through which data analysis can be drawn inferences from data sets without labelled responses.

G. Model Building

K-means is an iterative algorithm and does following important steps:

1) It starts with K as the input which is how many clusters you want to find. Place K centroids in random locations in your space.
2) Now, using the Euclidean distance between data points and centroids, assign each data point to the cluster which is close to it.
3) Recalculate the cluster centers as a mean of data points assigned to it.
4) Repeat 2 and 3 until no further changes occur.

Difference between k-means and k-mean++ is that k-means only changes how to initialize centroids. K-means++ is an algorithm for choosing the initial values for the k-menas clustering algorithm.

H. Search and Information Retrieval

The users can search information according input parameters such as skills, job type, industry type and location. Relevant information will be displayed with their needed information. Information retrieval is the activity to obtain information resources relevant to user information need from a collection of resources. Automated information retrieval systems are used to reduce information overload.

I. Evaluation for k-means and k-means++

a) Euclidean Distance and Elbow Method Evaluation Criteria:

Euclidean Distance is the sort of space where lines that start parallel stay parallel, and always stay the same distance from each other. One method to validate the no. of clusters is the elbow method. The idea of the elbow method is to run k-means and k-means++ clustering on the data set for a range of values of k (0-19) and for each value of k calculate the sum of distance between data and cluster center. The elbow

method is used to define the optimal number of clusters in algorithm. The lesser the sum of distance between data and cluster center, the greater the performance will be. As the graph shows that when the k increases, sum of distance between data and cluster center will also increase.

b) Silhouette Analysis

Silhouette method is used to interpret and validate of consistency within clusters of data. The clustering models are evaluated as Silhouette Score. Silhouette analysis measures distance between instances within the cluster and instances in the nearest cluster. This measure has a range of [-1, 1] where high value indicates that the objects is well matched to its own cluster and poorly matched to neighboring clusters.

The Silhouette Coefficient is calculated using following formula :

$$(b - a) / \max(a, b) \quad (4)$$

where

a: the mean intra-cluster distance

b: the mean nearest-cluster distance for each sample. [14]

Silhouette Evaluation Criteria:

TABLE I
SILHOUETTE EVALUATION CRITERIA

Score	1	-1	0
Performance	Best (data instance is close to the center of the cluster)	Worst	Overlapping clusters(on the border between two clusters)

Search Result Evaluation Discounted Cumulative Gain is a measure of ranking quality. It is used to measure effectiveness of search algorithms in information retrieval. Highly relevant jobs are shown in earlier in search result. DCG measures the usefulness, or gain, of a document based on its position in the result list. The gain is accumulated from the top of the result list to the bottom, with the gain of each result discounted at lower ranks. To select the best clustering approach, sum of squared distances of samples have been used to their closed cluster center.

$$DCG_p = \sum_{i=1}^{p} \frac{rel_i}{log_2(i+1)} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2(i+1)} \quad (4)$$

Where,

reli is the graded relevance of result at position i. [10]

So, CG is unaffected by changes in ordering of search results and hence, DCG is used for more accurate measure.

Discounted Cumulative Gain - Evaluation Criteria:

TABLE II
DCG EVALUATION CRITERIA

Rate	3	0
Search Result	Most Relevant	Least Relevant

IV. RESULTS AND DISCUSSION

A. System Description

The information retrieval system is created, thus helps applicants find the most relevant job post according to the user's query. The system analysis, comparison, implementation is done in this section. The experiment is done in Jupyter Notebook with in Mac pro operating system.

B. Tools and Technologies

The followings tools and technologies have been used in this research:

1) Python with Libraries: NumPy, pandas, matplotlib, scikit-learn
2) Miniconda
3) Jupyterlab

C. Experiment and Results

In general K-means++ is better than K-means. So, this research compares and justifies that the performance of K-means++ is better than K-means. The experiments shows the performance analysis of K-means and K-means++.

1) Elbow Method: **K-means Elbow Method**

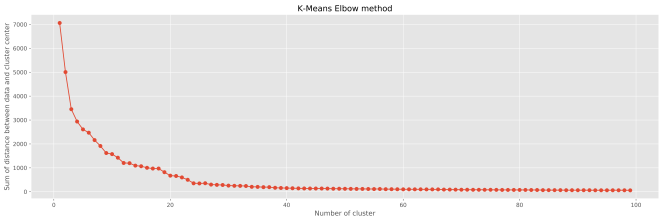

Fig. 7. K-means Elbow Method

TABLE III
K-MEANS ELBOW METHOD

Number of Cluster	20	40	60	80
Sum of distance between data and cluster center	672.30	152.04	101.59	75.38

The above figure is the experiment of K-means Elbow Method. 20 number of clusters are created for experiment. No. of clusters: 20, 40, 60 and 80 show sum of distance between data and cluster center : 672.30, 152.04, 101.59, and 75.38 respectively.

K-means++ Elbow Method

TABLE IV
K-MEANS++ ELBOW METHOD

Number of Cluster	20	40	60	80
Sum of distance between data and cluster center	396.65	122.18	71.33	51.41

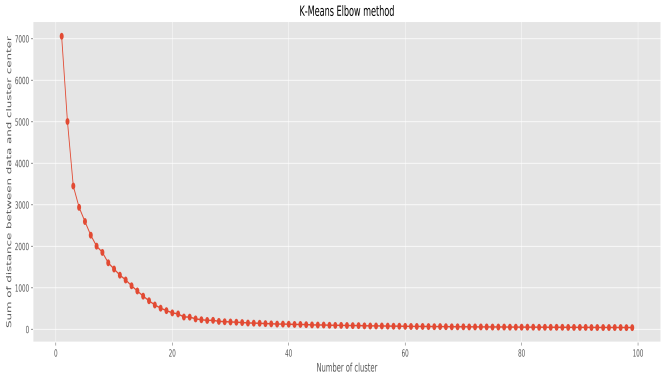

Fig. 8. K-means++ Elbow Method

The above figure represents the experiment of K-means++ and 20 number of clusters are created for system experiment. No. of clusters: 20, 40, 60 and 80 show sum of distance between data and cluster center : 396.65, 122.18, 71.33, and 51.41 respectively.

The complexity will increase if high no. of clusters are selected. So, 20 no. of cluster is selected for better model building.

Performance Comparison Elbow Method: K-means vs K-means++

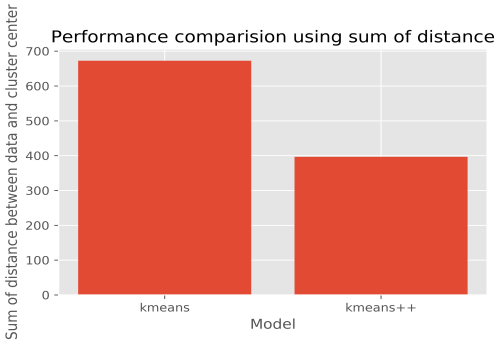

Fig. 9. Performance comparison Elbow Method

TABLE V
PERFORMANCE COMPARISON ELBOW METHOD

Algorithm	K-means	K-means++
Sum of distance between data and cluster center	672.32	396.65

- K-means has 672.31 sum of distance between data and cluster center
- K-means++ has 396.65 sum of distance between data and cluster center
- The lesser the distance, the greater the performance will be. So the K-means++ has higher performance.

D. Performance Comparison Silhouette Analysis

K-means has 0.64 and K-means++ has 0.70. K-means++ has higher silhouette coefficient than k-means. The best value is 1

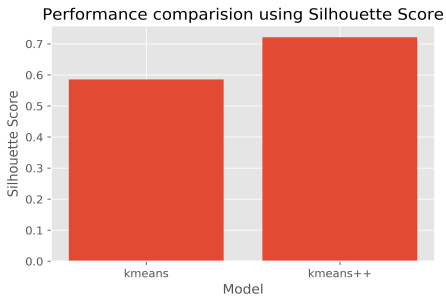

Fig. 10. Silhouette Analysis

TABLE VI
SILHOUETTE ANALYSIS

Algorithm	K-means	K-means++
Silhouette Score	0.5853345734051827	0.7213995183056551

and the worst value is -1. Values near 0 indicate overlapping clusters. Thus, K-means++ has value near to 1 than k-means so k-means++ has higher performance than k-means.

This research shows that K-means++ has higher performance according to the Elbow method and Silhouette Method. Hence, K-means++ is used for system implementation.

V. SEARCH EVALUATION

A. Search Result Evaluation

TABLE VII
SEARCH EVALUATION

Id	Skills	Cluster (predict label)
401	statistical software, Machine Learning, SAS, Python	11

TABLE VIII
NORMALIZATION SCORE CALCULATION

Skills	Frequency	Cluster Frequency	Score	Cluster (Predict Level)
Statistical Software	147	21	21/147=0.142	11
Machine Learning	2297	234	34/2297 =0.101	11
R	2234	206	206/2234 =0.0922	11
SAS	941	73	73/941 =0.077	11
Python	3325	251	251/3325 =0.075	11

Calculating normalized score for each skill that is assign to each cluster

Total Score = 0.142+0.101+0.0922+0.077+0.075 = 0.490006

B. *Discounted Cumulative Gain*

Subjective evaluation and DCG:
To find the quality of the search result, human rating is used for the search result and then aggregate the result using discounted cumulative gain.
DCG is used for accurate measure.

- It has a rate of the search result as 3 being the most relevant and 0 being the least relevant.
- Ideal relevance order = [3, 3, 3, 2, 2].

The following are the 4 queries and evaluating top 5 search:
Sample Query:

```
input_skills = [statistical software,
Machine Learning,SAS, Python]
input_state = [CA]
input_industry = [Internet and Software]
input_job_type = data_scientist
predict level = 11
Score = cluster frequency /  total frequency
% \end{lstlisting}
```

Now, 4 different queries and Top 5 result evaluation:

```
S [1] = [3, 3, 2, 2, 2]
S [2] = [3, 2, 2, 1, 1]
S [3] = [3, 1, 2, 1, 0]
S [4] = [3, 3, 3, 1, 0]
```

Average search performance using DCG formula : 0.805918405200146

VI. CONCLUSION AND FUTURE WORK

A. *Conclusion*

This research works on K-means++ clustering algorithm for system implementation. The experiment comparison between K-means and K-means++ is done in order to justify that K-means++ has higher work performance than K-means. K-means starts with allocating cluster centers randomly and then looks for better solutions. Whereas, K-means++ starts with allocation one cluster center randomly and then searches for other centers given the first one. So both algorithms use random initialization as a starting point, so can give different results on different runs. Since, K-means++ is better than K-means, this research tries to justify that K-means++ has higher performance than K-means. So, K-means++ has been chosen for the research work.

B. *Limitation*

This research works on information retrieval. The followings are the some of limitation of this research work:

- This research only works on IT jobs data sets from indeed website.
- This research only deals with information retrieval based on user query. It does not work on recommending jobs according to the user profile.

C. *Future Work*

Since, this research is about information retrieval, it displays the most relevant information. It is about finding the most relevant information from the collection of the jobs based on the user query. The automatic job recommendation can be implemented in future. If the applicant profile matches with jobs then it automatically suggest. Also, different data sets can be used for different results. Moreover, other clustering algorithm can be implemented to improve the performance.

REFERENCES

[1] D. Dwivedi, "https://towardsdatascience.com," 7 May 2018. [Online]. Available: https://towardsdatascience.com/machine-learning-for-beginners-d247a9420dab.
[2] A. I. M. I. Alvida Mustika Rukmi, "Using k-means++ algorithm for researchers clustering," Reserch Gate, 9 December 2017.
[3] S.R.N and Arunachalam2, "Movie ranking using k-means clustering algorithm in data mining," International Journal of Mathematics and Computer Applications Research (IJMCAR), vol. 3, no. 2, June 2013.
[4] M. M. a. M. A. Bashar Aubaidan, "Comparative study of K-means and K-means++ clustering algorithms on crime domain," Journal of Computer Science, 2014.
[5] Silvio Lattanzi, Christian Sohle, "A Better k-means++ Algorithm via Local Search," International Conference on Machine Learning, 2019.
[6] A. A. Pranita Sathe1, "Personalized Travel Recommendation Using K-Means Clustering," International Journal of Advanced Trends in Computer Applications, vol. 4, no. 1, pp. 6-9, January 2017.
[7] "https://sites.google.com," [Online]. Available: https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm
[8] Elroy, "Kaggle," October 2018. [Online]. Available: https://www.kaggle.com/elroyggj/indeed-dataset-data-scientistanalystengineer.
[9] Tim, "Stack Exchange," 1 January 2015. [Online]. Available: https://stats.stackexchange.com/questions/130888/k-means-vs-k-means.
[10] S. S, "https://machinelearningmedium.com,"[Online].Available: https://machinelearningmedium.com/2017/07/24/discounted-cumulative-gain/
[11] Fazal Rehman Shamil, "https://t4tutorials.com", [Online]. Available: https://t4tutorials.com/min-max-normalization-of-data-in-data-mining/
[12] A. Banes, "https://www.kaggle.com/," 2018. [Online]. Available: https://www.kaggle.com/antrellbanes/using-categorical-data-with-one-hot-encoding.
[13] F. R. Shamil, "https://t4tutorials.com/," 2019. [Online]. Available: https://t4tutorials.com/min-max-normalization-of-data-in-data-mining/.
[14] EduPristine, "https://www.edupristine.com," 21 July 2015. [Online]. Available: https://www.edupristine.com/blog/beyond-k-means.