Searching Suitable Employee for the Work by Using Partitioning Methods

Win Hay Mar Nwe, Nang Saing Moon Kham
University of Computer Studies, Yangon
winhaymarnwe.ucsy@gmail.com ;

Abstract

Conventional database query methods are inadequate to extract useful information from huge data banks. Cluster analysis is one of the major data analysis methods, and process of grouping a set of physical or abstract objects into classes of similar objects. A cluster is the collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. There are many approaches for clustering method. Partitioning is the well-known and efficient algorithm and K-Mean is the partitioning method and widely used in many applications. In this paper, K-means clustering method is used to find the appropriate employee from many job seekers and give that information to the employer to choose for their work. The k-means clustering methods are determining level of employee based on their profiles. Thus, many companies are easy to find appropriate employee for their work.

1. Introduction

Clustering is one of the fundamental operations in data mining. Clustering plays an outstanding role in data mining applications such as scientific exploration, information retrieval and text mining, spatial database applications, web analysis, marketing medical diagnostics, computational biology and many others. Due to its immense applications in various areas, clustering has been a highly active topic in data mining research.

Educational mismatch and skill mismatch raise concern among Employers and recruiter in the employee search areas. Human capital, job matching, and occupational mobility theories argue that educational mismatch is basically a transient situation due to either insufficient information exchange between employers and employees or to deficits in some areas of human capital. Relational queries cannot produce the required results for the employer. Clustering analysis is suitable to group the similar application forms and it is the most appropriate for the employers. Partitioning methods are widely used in Cluster analysis, and K-Means clustering algorithm is well known partitioning method. This paper presents the clustering of application forms in order to get the appropriate applications for the employer.

The rest of the paper is organized as follow: Section 2 is the related work. Data mining and clustering is presented in section 3. In section 4, K-Mean is described. Section 5 is the proposed system design and section 6 is the system implementation and sample case study for K-Mean clustering. Section 7 is the conclusion of the system.

2. Related Work

Clustering is important in many different fields such as data mining, image compression and information retrieval. [4] provided an extensive survey of various clustering techniques.

Clustering algorithms can be divided into hard and soft clustering algorithms. According to [5], there are four different kinds of clustering algorithms: hierarchical, partition, model fitting and density based. These algorithms form clusters by putting each item into a single cluster. Soft clustering allows each item to associate with multiple clusters, by introducing a membership function W_{ij} between each cluster-item pair to measure the degree of association. Expectation-maximization [2] served as the basis of many soft-clustering algorithms.

Many clustering techniques have been used for clustering. Willett [11] provided a survey on applying hierarchical clustering algorithms into clustering documents. Buckshot selects a small sample of documents to pre-cluster using a standard clustering algorithm and assigns the rest of the documents to the clusters formed. Fractionation splits the N documents into 'm' buckets where each bucket contains N/m documents. Fractionation takes an input parameter ρ, which indicates the reduction factor for each bucket [11]. The standard clustering algorithm is applied so that if there are 'n' documents in each bucket, they are clustered into n/ρ clusters. Now each of these clusters is treated as if they were individual documents and the whole process is repeated until 'K' clusters are left.

Most of the algorithms above use a word-based approach to find the similarity between two documents. In [12] a phrase-based approach called STC (suffix-tree clustering) was proposed. STC used a suffix-tree to form common phrases of documents enabling it to form clusters depending not only on individual words but also on the ordering of the words.

3. Data Mining

Data mining refers to extracting or mining knowledge from large amounts of data. It can be also defined as Knowledge Discovery in Databases, or KDD. Data mining is the process of extracting interesting information or patterns from large information repositories such as: relational database, data warehouses, XML repository, etc. Also data mining is known as one of the core processes of Knowledge Discovery in Database (KDD). The main process of KDD is the data mining process. In this process different algorithms are applied to produce hidden knowledge. Then, comes another process called post-processing, this evaluates the mining result according to users' requirements and domain knowledge. [1]

Various data mining techniques are applied to the data source; different knowledge comes out as the mining result. That knowledge is evaluated by certain rules, such as the domain knowledge or concepts.

3.1 Cluster Analysis

Clustering is a process of partitioning a set of data (or objects) in a set of meaningful sub-classes, called clusters. It can be defined as the process of organizing objects in a database into clusters/groups such that objects within the same cluster have a high degree of similarity, while objects belonging to different clusters have a high degree of dissimilarity. It is unsupervised classification and there is no predefined class. It helps users understand the natural grouping or structure in a data set. A cluster is a collection of data objects that are similar to one another and thus can be treated collectively as one group. [4]

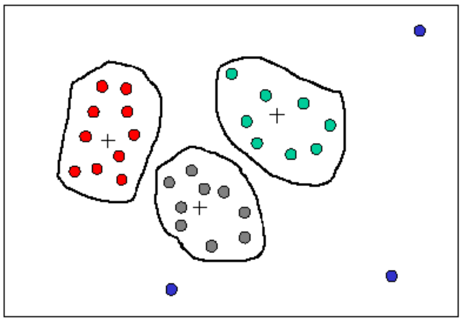

Figure 1: Sample Set of Cluster

Clustering process can be classified as follows:

Hierarchical Clustering: Hierarchical algorithms find successive clusters using previously established clusters. Hierarchical algorithms can be agglomerative (bottom-up) or divisive (top-down). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.

Partitioning Methods: Partitioning clustering algorithms typically determine all clusters at once, but can also be used as divisive algorithms in the hierarchical clustering.

Density-based clustering algorithms are devised to discover arbitrary-shaped clusters. In this approach, a cluster is regarded as a region in which the density of data objects exceeds a threshold. DBSCAN and OPTICS are two typical algorithms of this kind.

Two-way clustering, co-clustering or biclustering are clustering methods where not only the objects are clustered but also the features of the objects, i.e., if the data is represented in a data matrix, the rows and columns are clustered simultaneously.

4. Clustering by Partitioning with K-Mean

A common clustering algorithm is k-means clustering algorithm. It is distance based, unsupervised and partition. It is the simplest and most commonly used clustering algorithm, especially with large data sets. K-means clustering increases the accuracy of the system. Distance-based clustering involves determining a distance measure between pairs of data objects, and then grouping similar objects together into clusters.

The k-means algorithm uses the mean value of the objects in a cluster as the cluster. It is built upon four basic operations (1) selection of the initial k means for k clusters, (2) calculation of the dissimilarity between an object and the mean of a cluster, (3) allocation of an object to the cluster whose mean is nearest to the object, (4) re-calculation of the mean of a cluster from the objects allocated to it so that the intra cluster dissimilarity is minimized. It involves following steps: [10]

- Define a set of items (n-by-p data matrix) to be clustered.
- Define a chosen number of clusters (*k*).
- Randomly assign a number of items to each cluster.

The *k*-means clustering repeatedly performs the following until convergence is achieved:

- Calculate the mean vector for all items in each cluster.

- Reassign the items to the cluster whose center is closest to the item.

In statistics and machine learning, k-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. K-Mean algorithm is described as follows:

Algorithm K-Mean Clustering

Input: The number of cluster k and a database containing n objects.

Output: A set of k clusters that minimizes the squared – error criterion.

Method:

```
    arbitrarily choose k objects as the initial cluster centers;
    repeat
    {
    (re) assign each object to the cluster to which the objects is the most similar, based on the mean value of the objects in the cluster;
    Update the cluster means, i.e., calculate the mean value of the objects for each cluster;
    }
    Until no change;
```

4.1 Cosine Similarity Algorithm

Cosine Similarity Algorithm is used to compute the (distance) Similarity between user application forms. It measures the relevant degree between user applications. In this system, user applications are prepared as term vectors and cosine coefficient can be computed for those vectors. Equation for computing Cosine similarity is as follows:

$$\mathrm{Sim}\ (a, b) = \cos\theta = \frac{a \bullet b}{|\,a\,|\ x\ |\,b\,|}$$

$$\mathrm{Sim}\ (S_a, S_b) = \frac{\sum_{i=1}^{n}(S_{ai} \bullet S_{bi})}{\sqrt{\sum_{i=1}^{n} S^2_{ai} * \sum_{i=1}^{n} S^2_{bi}}}$$

Eqn. (1)

Where, a = attributes of first employee

b = attributes of second employee

n = number of total words in attributes of first and second employee.

S_{ai} = weight of term i for first employee (a)

S_{bi} = weight of term i for first employee (b)

Advantages of Cosine Similarity Algorithm

It is simple, mathematically based approach, providing partial matching and ranked results. Cosine similarity works well in practice. It calculates the similarity values in a precise way and expresses the differences in terms of bits of information. It allows efficient implementation for large document collections

5. Proposed System

This paper presents the grouping of similar application forms (CV, resume) so that employers are easy to find the employee he / she wants.

K-Means clustering algorithm is used to group the similar application forms. It is the partitioning algorithm and widely used in many real world applications. In this system, applicants (employee) submit their application forms to the job search site, and the recruiter groups the similar application forms and redirect to employers. They select the most appropriate application forms to sit the interview. Instead of reviewing all application forms, reviewing selected group saves the employer's time and more effective to get the required application forms.

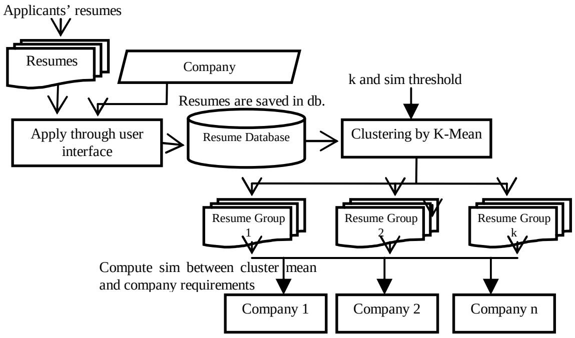

Figure 2: Proposed System overview

Figure 2 presents the overview of the system. According to the above figure, when applicant submits the application form, it will go to Resume database at the recruiter site (job search site). There, application forms are clustered by K-Means clustering algorithm, which groups the similar resumes into the same group. In this system, k and similarity threshold are provided by user. Employee application with high similarity values are clustered into same group.

Then groups of application forms are redirected into the corresponding company (employer) for reviewing, this process is computed by finding similarity between company requirements and cluster mean. K-Mean clustering algorithm has been described in Section 4. In the similarity computation, Cosine Similarity (Eqn. 1), is used.

6. System Implementation

This system is implemented using ASP .Net C#. It is in the form of web-based job recruitment system, accepting employee's application forms (resumes) and then redirected to appropriate company based on company's requirements. Class diagram for this proposed system is shown in Figure 3.

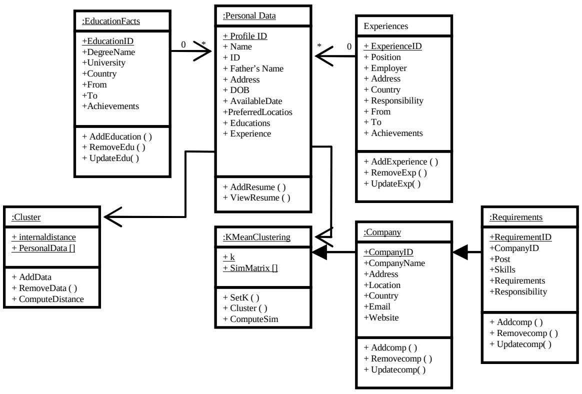

Figure 3: Class Diagram for clustering Employee Application forms

6.1. Experimental Result

In the implementation, employees who are searching for new jobs have to submit their application forms into the system. Those application forms are saved into the database. They are redirected into related company, that is searching new applicants for their requirements. Before redirecting application forms, they are grouped into clusters. Application forms with similar attributes are clustered into same group. In this system, similarity between employee's attributes is computed based on education, experiences, skills using Cosine similarity algorithm.

In the case study of clustering employee application forms, example data sets shown in following tables are used. Sample Employee application database is shown in Table 1.

Table 1: Sample Employee Application Forms

ID	Name	Skills	Education	Experience
E1	Ma Win Hay Mar Nwe	Java,C#,ASP .Net,PHP	B.C.Sc; M.C.Sc	Programmer
E2	Ma Su Mon	Microsoft Word,Office Word,Power Point	B.Sc	Office Staff
E3	Mg Kaung Kaung		B.Sc	
E4	Mg Aung Myat	LCCI (Level-2)	B.Sc	Staff
E5	Mg Htun Htun	English Language	B.A	
E6	Ma Yin Yin	Java	B.C.Sc	Programmer
E7	Ma Win May	C#,Java Basic	M.C.Sc	Programmer

In order to compute in vector space model, terms are extracted from employee data (skills, education and experience).
For example, Terms of Employee ID = E1
"java", "c#", ".net", "asp", "php", "b.c.sc", "m.c.sc", "programmer"
Terms of Employee ID = E2
"microsoft word", "office word", "power point", "b.sc", "office staff"
Terms of Employee ID = E3
"b.sc"
Those terms are prepared as vector and applied to Cosine Similarity algorithm.

Table 2 is the similarity matrix (computed by cosine similarity algorithm) for resumes in table 1.

Table 2: Similarity Matrix

	E1	**E2**	**E3**	**E4**	**E5**	**E6**	**E7**
E1	1	0	0	0	0	0.521	0.544
E2	0	1	0.333	0.569	0	0	0
E3	0	0.333	1	0.333	0	0	0
E4	0	0.569	0.333	1	0	0	0
E5	0	0	0	0	1	0	0
E6	0.521	0	0	0	0	1	0.526
E7	0.544	0	0	0	0	0.526	1

When k is set to 2 with similarity threshold = 0.2. Clusters as in Table 3 are generated.

Table 3: Generated Clusters

Cluster 1				
ID	**Name**	**Skills**	**Education**	**Experience**
E1	Ma Win Hay Mar Nwe	Java,C# , ASP. Net, PHP	B.C.Sc, M.C.Sc	Programmer
E6	Ma Yin Yin	Java	B.C.Sc	Programmer
E7	Ma Win May	C#, Java Basic	M.C.Sc	Programmer
Cluster 2				
ID	**Name**	**Skills**	**Education**	**Experience**

E2	Ma Su Mon	Microsoft Word, Office Word, Power Point	B.Sc	Office Staff
E3	Mg Kaung Kaung		B.Sc	
E4	Mg Aung Myat	LCCI (Level-2)	B.Sc	Staff

Clustering into similar groups help employers to find the required employee easily, finding in the small groups instead of finding in large amount data.

7. Conclusion

Nowadays, there are a lot of job seekers who submit online resume into job search engine. It makes a large bunch of data and making difficult for the employers to get the appropriate applicant (employee). This system presents grouping the similar application forms by using K-Means clustering algorithm. It saves times for employers and effective in finding the appropriate employees. K-Means clustering is well-known clustering algorithm and it is suitable for grouping similar data from large datasets.

8. References

[1] Chen, M.S., Han, J., and Yu, P.S., "Data Mining: An Overview from a Database Perspective", IEEE Transactions on Knowledge and Data Engineering, 8(6): 866-883, 1996.

[2] Dempster, A.P., Laird, N.M., and Rubin, D.B., "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, Series B, 39(1), 1-38, 1977.

[3] Hill, D.R., "A vector clustering technique, in: Samuelson (Ed.), Mechanized Information Storage, Retrieval and Dissemination", North-Holland, Amsterdam, 1968.

[4] Jain, A.K., Murty, M.N., and Flynn, P.J., "Data Clustering: A Review, ACM Computing Surveys". 31(3): 264-323, Sept 1999.

[5] Krzanowski, W.J., and Marriott, F.H., "Multivariate Analysis: Classification, Covariance Structures and Repeated Measurements". Arnold, London, 1998.

[6] Moore, J., Han, E., Boley, D., Gini, M., Gross, R., Hastings, K., Karypis, G., Kumar, V., and Mobasher, B., "Web Page Categorization and Feature Selection Using Association Rule and Principal Component Clustering", In Proceedings of seventh Workshop on Information Technologies and Systems (WITS'97), December 1997.

[7] Munoz, A., "Compound key word generation from document databases using a Hierarchical clustering ART Model, Intelligent Data Analysis", Jan 1997. http://www-east.elsevier.com/ida/browse/96-5/ida96-5.htm

[8] Murty, M.N., and Jain, A.K., "Knowledge-based clustering scheme for collection management and retrieval of library books, Pattern recognition", 28, 946-964, 1995.

[9] Nigam, K., Mccallum, A.K., Thrun, S., and Mitchell, T., "Text Classification from Labeled and Unlabeled Documents using EM. Machine Learning", 39(2-3):103-134, 2000.

[10] Roussinov, D., Tolle, K., Ramsey, M., and Chen, H., "Interactive Internet search through Automatic clustering: an empirical study", In Proceedings of the International ACM SIGIR Conference, pages 289-290, 1999.

[11] Willett, P., "Recent trends in hierarchical document clustering: a critical review, Information processing and management", 24: 577-97, 1988.

[12] Zamir, O., and Etzioni, O., "Web document clustering: a feasibility demonstration, in Proceedings of 19th international ACM SIGIR conference on research and development in information retrieval (SIGIR 98)", 1998, pp 46-54.