EJASA:DSS
Electronic Journal of Applied Statistical Analysis:
Decision Support Systems and Services Evaluation

**EJASA:DSS – Electronic Journal of Applied Statistical Analysis: Decision Support Systems and Services Evaluation**
http://siba-ese.unisalento.it/index.php/ejasa_dss/index
e-ISSN: 2037-3627
DOI: 10.1285/i2037-3627v5n1p57

# Implementation of a Training Courses Recommender System based on k-means algorithm

By Mohammad, Alhaidey

Published: 29 December 2014

*Electronic Journal of Applied Statistical Analysis: Decision Support Systems and Services Evaluation*
Vol. 05, Issue 01, December 2014, 57-66
DOI: 10.1285/i2037-3627v5n1p57

# Implementation of a Training Courses Recommender System based on k-means algorithm

Heba Mohammad*[a] and Hana Alhaidey[b]

[a] *College of Computer and Information Science,, Al Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia*

Published: 29 December 2014

Providing the right professional training courses for employees is a critical issue for organizations as well as employees. This paper aims to propose a framework for building a recommender system for training courses within educational institutions in Saudi Arabia. The paper shows the result of implementing the early two phases of the proposed framework. These two phases are based on collecting data and applying k-mean clustering. The results show that 6 clusters had been generated.

## 1 Introduction

In the 21st century, information and communication technologies are evolving. We are transferring from e- commerce to u- commerce. Where data and information had been accumulated rapidly in huge amounts within Organizations repositories. Thus, executives are striving to reap the most benefits from massive data they have. In order to enjoy competitive advantages on their market. A study by MIT published on its website on 2010, shows that best performers in market where those organizations who own the analytical capabilities and interpretation skills in the era of Big Data. Thus, Organization are seeking now to build their analytical capabilities to take in their considerations the wide ranges of decisions various paths that could be taken by them.

Hence, organizations are relying on different techniques and approaches in order to extract the hidden patterns and paths. This tends to the use of certain information technologies to help in discovering, sharing, incorporating and distributing valuable knowledge and information. One of the most efficient used techniques is data mining (Alex

*Corresponding author: hkmohammad@ccis.imamu.edu.sa

ISSN: 2037-3627
http://siba-ese.unisalento.it/index.php/ejasadss/index

et al., 2000).

The main concern of data mining is the process of extracting helpful knowledge from huge databases using various tools. In practice, data mining is helpful for managing organization data and knowledge in two manners; sharing common business intelligence, and using data mining as a technique for extending human knowledge. Therefore, data mining techniques can assist organizations in finding hidden knowledge in huge databases (Silwattananusarn and Tuamsuk, 2012).

Recently, due to large amounts of available contents and items, various recommender systems have been constructed to be used in suggesting the most appropriate contents for users based on creating a list of personalized predictions for users. The purpose of these systems is the reverse of the traditional information retrieval process where these systems can demonstrate the interesting contents of users depending on their historical behavior instead of answering a certain query. These systems have become very common in the last few years and applied in several applications, such as in suggesting and predicting the most important books, movies, social tags, products, news and articles for various types of users. Furthermore, these systems have been proposed for financial services, and twitter followers and live insurances (Lathia et al., 2008; Gupta et al., 2013).

The main benefits of using recommender systems are helping users in their decision-making process to decide the most optimal contents for them, helping users in discovering interesting items that they do not know and helping businesses in creating more income through modified discussion marketing that aims to help each user individually. In other words, recommender systems can be defined as systems that offer a set of predictions, suggestions or opinions to help users in assessing and choosing items. Various companies, such as small and medium enterprises, to train their employees in order to enhance their abilities and assist them to perform their works more effectively, also use recommender systems. Furthermore, these systems help companies in ensuring that all their employees have up-to-date knowledge and skills (Darzi et al., 2010). Thus, this work aims to propose a framework for building a training course recommender system based on K-mean algorithms.

## 2 The use of Data Mining in Predicting the Performance of Employees and Offering improvement Suggestions

In the recent years, various researchers conducted works using data mining techniques to extract rules and predict specific behaviors in various fields, such as education, science, human resources medicine and biology. In the educational field, Al-Radaideh et al. (2006) proposed the use of data mining technique in predicting the performance of students in a certain university. Moreover, the proposed work by Schwab (1991) focused on recognizing on describing university teachers performance based on the number of published or cited researches. Beikzadeh et al. (2008) developed a data mining based system to predict improvements and performances for higher educational institutions.

The main concern of various industries and organizations is predicting and evaluating the performance of their employees. Generally, the measurement of performance depends on using the generated units by employees in their job in a certain time period. However, the employees performance measurements differ based on the context. Various organizations have spent significant resources and time to construct measurement systems for measuring the performance of their employees. Richason proposed that the main stages of measuring the employee performance are: establishing a baseline, identifying and subtracting performance limiters, creating a spreadsheet and re-evaluating the performance measurement. Currently, data mining techniques are widely used for that purpose in various organizations.

Chen et al. (2007) proposed the use of data mining technique for developing a web-based Employee Training Expert System (ETES) to infer the appropriate learning type for employees. The developed system depended on using association rule mining for discovering and suggesting the optimal learning map and training strategies for personal learning. This system has offered various training courses for employees depending on their learning skills, careers and records.

Jantan et al. (2010a) used data mining classification technique to predict the performance of employees in a certain organization using the past performance records in datasets. The produced classification rules can be utilized in predicting the potential talent of certain tasks in that organization. Furthermore, various tasks can be accomplished using this approach, such as choosing new staff, predicting the current and future performance of employees, and setting training needs for employees. Gupta et al. (2013) used data mining technique to predict employees performance in IT industries based on their previous experience knowledge through past performance assessment data. In the proposed method, patterns are initially identified and produced using data mining to define and classify employees who have similar characteristics and relate profiles of employees for the most suitable projects.

## 3 Attributes that influence Performance of Employees

Generally, there are various factors that affect on the performance of employees, such as personal, professional and educational factors. Chien and Chen (2008) developed a data mining based model for enhancing the selection of employees and predicting the behavior of newly interviewees. The predicted performance can be a base for decision makers, which help them in taking decisions concerning employing those interviewees or not. The main used attributes in this research were gender, age, experience, marital status, school tire, education level and taken courses. After that, they excluded three of them; gender, age and marital status. Thus, no discrimination can be presented in the personal chosen process. Researchers explored that the performance of employees is highly influenced by three attributes; job experience, school tire and degree.

Kahya (2007) offered a study to determine the main attributes that influence the job performance. They studied the influence of job satisfaction, education, income, working conditions and experience on the job performance. They found that the employees work positions have a major positive effect on their performances, while both the working environment and conditions can have negative and positive effects on the performance. In other words, highly qualified and educated employees demonstrated dissatisfaction of terrible working conditions, which in turn have influenced on their performance, while low qualified and educated employees demonstrated satisfaction concerning the bad working conditions. Furthermore, it was found that the experience has a positive impact on the performance, while the education has not a clear impact on it.

Jantan et al. (2010a) proposed an HR system design to predict the talent of applicant depending on both the previous experience and filled information in the HR application with the use of data mining techniques. They focused on studying the effects of certain attributes; social responsibility, training and professional qualification. Various hybrid data mining techniques have been applied to discover the main prediction rules; Decision Tree (DT), Artificial Neural Network (ANN) and Rough Set Theory (RST). Those authors proposed another study that aims to suggest suitable human talent in Human Resource Management (HRM). The DT technique was used in the proposed study based on producing categorization rules for past HR records and evaluating them on hidden data to compute the precision. Furthermore, the generated rules were used to generate a system for predicting both the behavior and potential promotions of employees.

Chen and Gong (2013) used a data mining decision tree algorithm to analyze the engagement of IT employee. A model using this technique was designed to discover deep applicable engagement knowledge, where extracting rules from that model can help administrators in their decision-making process and this in turn can enhance their employees engagement.

Chen and Gong (2013) proposed that the main factors that can affect on the engagement analysis are the character and satisfaction of employees. The employees character includes the rank, working time, education background and age, while the employee satisfaction includes the satisfaction of staff, work, opportunity, life quality, salary and work procedure. Researchers agreed that the results and deep knowledge could provide basis for IT enterprises to develop flexible and targeted polices for employee engagement improvement.

## 4 Application of Data Mining in Recommender Systems

Data mining is a wide spectrum of mathematical software tools and modeling methods, which are utilized to discover patterns in data to be used in constructing models. In recommender systems, data mining is applied due to its ability to explain the used analysis techniques in inferring recommendation rules and construct enhanced recommendation models from huge databases. Data mining algorithms are used in recommendation systems to preprocess, analyze and classify data (Al-Radaideh and Al Nagi, 2012; Jantan et al., 2010b; Kanokwan et al., 2009; Van Meteren and Van Someren, 2000). Thus, data mining based recommender systems can produce their recommendations with the use of learned knowledge from users attributes.

Van Meteren and Van Someren (2000) proposed that content dependent recommender systems depend on using a dataset that contains the attributes and features that characterize each item, such as meta-data, keywords and descriptions. Sarwar et al. (2000) proposed that a collaborative-based recommender system includes three main stages; representation, formation of neighbourhood and creation of recommendations.

Furthermore, Jantan et al. (2010a) developed a data mining based system for developing the performance of employees and recommend the probable categorization techniques (nearest neighbour, decision tree and neural network) for their future performance. In the developed system, the patterns of employees performance can be detected from existing datasets in order to be used for predicting their future performance. In addition, Kanokwan et al. (2009) proposed a framework to develop an intelligent recommender system based on different data mining techniques such as clustering ( K-mean and two-step clustering techniques), association rules and fuzzy logic algorithm. Their proposed framework had been used to develop course recommender system for university students in Thailand.

Also, Al-Radaideh and Al Nagi (2012) proposed the development of a data mining based classification model for predicting employees performance. The adopted methodology to construct the proposed model was the Cross Industry Standard Process for Data Mining (CRISP-DM), which includes five stages: business recognition, data recognition, data training, construction and use. Furthermore, the used data mining technique was the Decision Tree (DT), in which various categorization rules were produced. This technique was used since there is no need for any domain expert knowledge as well as it is suitable for discovering exploratory knowledge and can offer a model with rules, which are human understandable and interpretable.

Other researchers suggest that recommendations systems could use hybrid algorithms in order to enhance the results (Blanco-Fernández et al., 2008; Gao et al., 2008; Mobasher et al., 2007). For example, Mobasher et al. (2007) proposed that using a combination of algorithms (hybrid) rather than single algorithm could give more enhanced results in generating recommendations and solve various issues that can be solved using traditional recommender systems. As an example, to solve the problems that generated from adding a new item, user dependent data can be joined with the content dependent data. Therefore, recommender systems are able to generate recommendations based on the similarity among users or contents. Hybrid recommender systems have been used to switch among various recommendation approaches depending on the location of users in a website. Blanco-Fernández et al. (2008) explored that the collaborative-based recommender system has an insufficient problem when there are inadequate users with the same profiles needed to produce significant suggestions. Conversely, the content-based recommender system has an overspecialization problem during offering similar items to those that are previously known by users. Thus, hybrid recommender systems are used to solve the problems of both the collaborative-based and content-based recommender systems based on using data mining techniques for conducting and filtering results for users. Gao et al. (2008) proposed that the most advanced hybrid technique is the semantic one. Thus semantic based recommender systems can address the restriction of previous recommender algorithms. This technique combines the semantic knowledge and the performances of their processes depending on a knowledge base that includes relationships among concepts.

## 5 Proposed Training Course Recommendation System Architecture

The proposed system supposed to recommend training courses for employees based on their previous records on the organization ( in our case is Nourah University). Employee information will be extracted from the database. The data will be subjected to preprocessing, cleaning and analyzed. The data is analyzed using different data mining techniques (k-mean clustering and association rules algorithm). Then the result from the analysis will be evaluated and weighted by prediction model to refine the recommendation. The final model will be integrated to online application system and used by the employees for testing. The online system will enable the employee to decide regarding the suggested training courses. Below is the proposed architecture of the system.

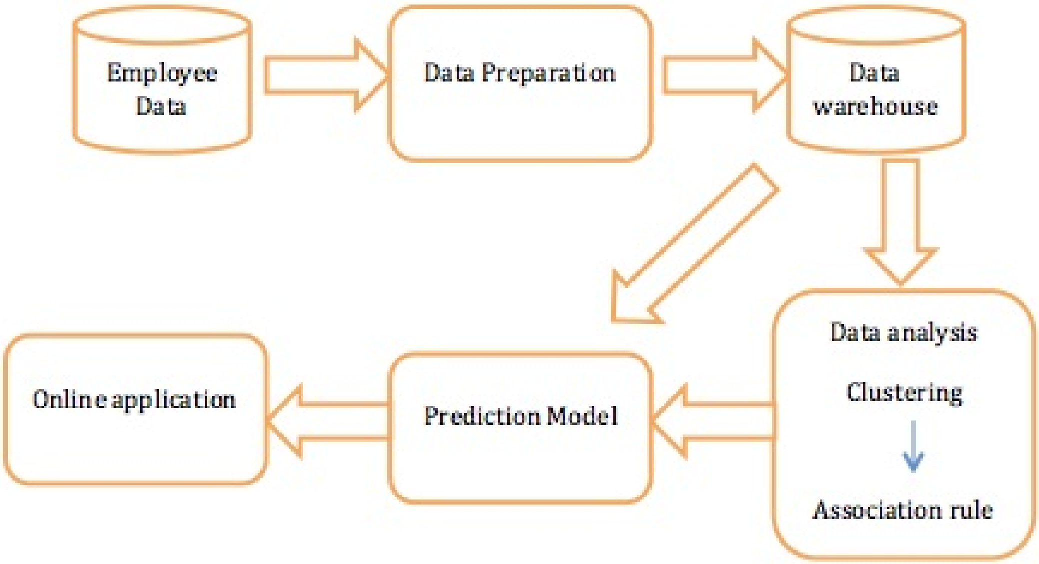

Figure 1: Proposed Recommendation System Architecture

## 6 Materials and Methods

This paper is concerned on the first two phases of the proposed recommendation system (data preparation and analysis) proposed recommendation system. Initially, the data concerning 300 employees in Princess Norah University were collected from the university Human Resource (HR) and then saved in a database. All information that might identify the employee had been removed due to privacy issues. Aslo, missing data had been handled or removed (Wong et al., 2004). The selected attributes were: education information which concerns the general specialization and taken training courses and professional information which concerns evaluation of employee performance, promotion probability and current academic rank. There were around 36 courses that are given for the female employees at Norah University.

Table 1: Attribute description

| | |
|---|---|
| Employee level | Represents the academic rank |
| Evaluation | Represents employee evaluation |
| Promotion | Represents if the employee had been promoted |
| Course | Training courses which had been taken by the employee |

The collected data about Princess Norah Universitys employees are then clustered using the k-simple mean clustering technique in the Waikato Environment for Knowledge Analysis (WEKA) software program 3.6.

## 7 Applying Data Mining: Simple K-means Clustering Technique

Clustering process is applied on database to divide the data into several significant subsets, which called clusters. The objects of each cluster have a common trait. The cluster analysis is conducted to find out a new group of categories. The clustering process represents finding the similarity among objects using the distance measure as shown in the following figure. Objects located in the same cluster must be close for each other. Therefore, for similar objects, the distance measure is short. The distance measure can be found in various ways, such as the Euclidean and Manhattan distances (Bhatia, 2004), . K-means clustering is a simple data mining technique, which used for cluster analysis to divide observations into $k$ clusters, where each observation belongs to that cluster with the closest mean. For a set of observations$x_1, x_2, \ldots, x_n$, in which each one is a d-dimensional real vector, the K-means clustering technique divides those observations into $k$ clusters; $K \preceq N : S = S_1, S_2, ., S_K$. The within-cluster sum of squares can be minimized using the following formula:

$$argmin_s = \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2$$

Where $\mu_i$ represents the mean of points in the cluster. To apply the k-mean clustering algorithm, the number of clusters ($k$) must be initially determined and the initial centroid of those clusters must be assumed. After that, the K-means clustering algorithm performs the following steps till reaching convergence and the centroids of groups are no longer in move (Jain et al., 2010):

- Determining the coordinates of the assumed centroids
- Determining the distance among each object and the centroids
- Grouping objects depending on the minimum distance (finding the closest centroid for the object

The most common utilized distance in the clustering process is the Euclidean distance, which studies the root of square differences among the coordinates of two objects. It can be expressed as follows:

$$d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ij} - x_{jk})^2}$$

## 8 Results and Discussion

The collected data were clustered using the K-means clustering technique in WEKA, where the selected number of clusters was 6.

Table 2: Results

| Cluster | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Number of Instances | 43 | 20 | 35 | 58 | 98 | 46 |
| Overall performance | 14% | 7% | 12% | 19% | 33% | 15% |

## 9 Conclusion

This paper proposed a framework to develop a recommender system based on different data mining techniques. However, here we presented data regarding the first two phases of the system architecture. Next we can apply the results to understand the associations between different attributes and clusters. However, it is widely known that clustering approach produce less-personal recommendations (Amatriain et al., 2011) and worse accuracy performance. Thus, Hybrid data mining technique could be used to improve the accuracy and provide more personalized recommendations. Also, we could use a larger dataset with more attributes in order to improve and enhance the recommendation process.

## References

Al-Radaideh, Q. A. and Al Nagi, E. (2012). Using data mining techniques to build a classification model for predicting employees performance. *International Journal of Advanced Computer Science and Applications*, 3(2).

Al-Radaideh, Q. A., Al-Shawakfa, E. M., and Al-Najjar, M. I. (2006). Mining student data using decision trees. In *International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan.*

Alex, B., Stephen, S., and Kurt, T. (2000). Building data mining applications for crm. *New York (etc.): McGraw-Hill.*

Amatriain, X., Jaimes, A., Oliver, N., and Pujol, J. M. (2011). Data mining methods for recommender systems. In *Recommender Systems Handbook*, pages 39–71. Springer.

Beikzadeh, M. R., Phon-Amnuaisuk, S., and Delavari, N. (2008). Data mining application in higher learning institutions. *Informatics in Education-An International Journal*, (Vol 7_1):31–54.

Bhatia, S. K. (2004). Adaptive k-means clustering. In *FLAIRS Conference*, pages 695–699.

Blanco-Fernández, Y., Pazos-Arias, J. J., Gil-Solla, A., Ramos-Cabrer, M., López-Nores, M., García-Duque, J., Fernández-Vilas, A., Díaz-Redondo, R. P., and Bermejo-Muñoz, J. (2008). A flexible semantic inference methodology to reason about user preferences in knowledge-based recommender systems. *Knowledge-Based Systems*, 21(4):305–320.

Chen, K.-K., Chen, M.-Y., Wu, H.-J., and Lee, Y.-L. (2007). Constructing a web-based employee training expert system with data mining approach. In *E-Commerce Technology and the 4th IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services, 2007. CEC/EEE 2007. The 9th IEEE International Conference on*, pages 659–664. IEEE.

Chen, Q. and Gong, Z. (2013). Data mining modeling of employee engagement for it enterprises based on decision tree algorithm. In *Information Management, Innovation Management and Industrial Engineering (ICIII), 2013 6th International Conference on*, volume 2, pages 305–308. IEEE.

Chien, C.-F. and Chen, L.-F. (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems with applications*, 34(1):280–290.

Darzi, M., Manesh, Z., Liaei, A. A., Hosseini, M., and Asghari, H. (2010). Fcrs: A fuzzy case-based recommender system for smes. In *Educational and Information Technology (ICEIT), 2010 International Conference on*, volume 3, pages V3–21. IEEE.

Gao, Q., Yan, J., and Liu, M. (2008). A semantic approach to recommendation system based on user ontology and spreading activation model. In *Network and Parallel Computing, 2008. NPC 2008. IFIP International Conference on*, pages 488–492. IEEE.

Gupta, P., Goel, A., Lin, J., Sharma, A., Wang, D., and Zadeh, R. (2013). Wtf: The who to follow service at twitter. In *Proceedings of the 22nd international conference on World Wide Web*, pages 505–514. International World Wide Web Conferences Steering

Committee.

Jain, S., Aalam, M. A., and Doja, M. (2010). K-means clustering using weka interface. In *Proceedings of the 4th National Conference*.

Jantan, H., Hamdan, A., Othman, Z., and Puteh, M. (2010a). Applying data mining classification techniques for employee's performance prediction. In *Knowledge Management 5th International Conference (KMICe2010)*, pages 645–652.

Jantan, H., Hamdan, A. R., and Othman, Z. A. (2010b). Data mining classification techniques for human talent forecasting. *Science And Technology*, pages 1–14.

Kahya, E. (2007). The effects of job characteristics and working conditions on job performance. *International journal of industrial ergonomics*, 37(6):515–523.

Kanokwan, K., Chun, C., and Wudhijaya, P. (2009). An intelligent recommendation system for student relationship management. In *The 8th International Conference on e-Business (iNCEB 2009) October28th-30th*.

Lathia, N., Hailes, S., and Capra, L. (2008). Trust-based collaborative filtering. In *Trust Management II*, pages 119–134. Springer.

Mobasher, B., Burke, R., Bhaumik, R., and Sandvig, J. J. (2007). Attacks and remedies in collaborative recommendation. *Intelligent Systems, IEEE*, 22(3):56–63.

Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2000). Analysis of recommendation algorithms for e-commerce. In *Proceedings of the 2nd ACM conference on Electronic commerce*, pages 158–167. ACM.

Schwab, D. P. (1991). Contextual variables in employee performance-turnover relationships. *Academy of Management Journal*, 34(4):966–975.

Silwattananusarn, T. and Tuamsuk, K. (2012). Data mining and its applications for knowledge management: A literature review from 2007 to 2012. *arXiv preprint arXiv:1210.2872*.

Van Meteren, R. and Van Someren, M. (2000). Using content-based filtering for recommendation. In *Proceedings of the Machine Learning in the New Information Age: MLnet/ECML2000 Workshop*.

Wong, K. W., Fung, C. C., Gedeon, T., and Chai, D. (2004). Intelligent data mining and personalisation for customer relationship management. In *Control, Automation, Robotics and Vision Conference, 2004. ICARCV 2004 8th*, volume 3, pages 1796–1801. IEEE.