A Novel Architecture for Content Recommendation in E-learning Environments Based on K-Means Clustering & Association Rule Mining

Ahmad A. Kardan, Nahid Ghassabzadeh Saryazdi, Hamed Mirashk

Advanced E-Learning Technology Laboratory
Department of Computer Engineering and Information Technology
Amirkabir University of Technology, Tehran, Iran
E-mail: {aakardan, nghassabzadeh, hamedmirashk}@aut.ac.ir

Abstract
Grouping e-learners based on their model in the e-learning environment is a key issue to build a personalized learning system. Recommender Systems can be useful to recommend learning resources or any other supportive advices to the learners. These systems could be used to suggest the contents being interested for learners in an e-learning environment. Different kind of algorithms such as user-based and item-based collaborative filtering have been used to establish a recommender system. In this paper, an innovative architecture for a recommender system (AELTRec) dedicated to the e-learning environments is introduced. This architecture simultaneously takes advantages of K-Means clustering technique and association rule mining. We first build a learner model based on PAPI learner model, which is the basis of learner grouping. Furthermore, K-Means is used to cluster the e-learner types. When groups of related interests have been established, the association rule mining techniques will be used to elicit the rules of the best content for each learner. Based on e-Learner groups, users can obtain content recommendation from the group’s opinions. It was expected that the proposed architecture has excellent performance.

Keywords: Personalization, Recommender Systems, Clustering Techniques, Association Rule Mining, E-learning

Introduction
In e-learning environments, presenting appropriate contents for each section of a course is an effective way for personalization of learning objects and improvement of learning. Therefore, recommendation of learning contents suitable for each section of a specific course has a significant role in improvement of learner’s functionalities.

Motivation and Necessity
The main goal of suggesting this architecture is an implication of a recommender system to recommend suitable learning resources for each section of a course.

It is obvious that exploiting this system in e-learning environment has major effect on improving learner’s performance. Because, if it is possible to help learners to utilize appropriate learning materials in its learning process it can improve learning progress in this process.

The Paper Content
Moreover, in this paper, at the very beginning in section 2, related works regarding to paper’s title are briefly discussed. After that, in section 3, the basic concepts such as clustering, association rules mining and recommender system are reviewed. Following that, in section 4, the proposing
architecture with detail description of the components and their functionalities are offered. At last, in fifth section, we summarize the obtained results of this research and consider future works which are appropriate in this area.

**Related works**

Offering a solution related to e-learning is different from other domains. (The most domains which are studied in recommender systems are presenting recommendations related to movies.) Special issues for a recommender system in e-learning are as follows (Zhang Kun et al, 2007):

- Items where are interested by learners, might not be usable for them. In compare with other conditions, suggestions are made according to learners’ preferences.
- Personalization should not be solely related to select learning items; rather it should consider their deliveries.

In overall, the general work process in recommender systems is as first to retrieve a series of entered data and information. Hence using different techniques are processed and finally results, which are the recommendations, are shown.

In Table 1, these three phases are analyzed and compared with the proposed method in this paper.

### Table 3. Comparing three main phases in recommender systems with the proposed method

<table>
<thead>
<tr>
<th></th>
<th>Information Receipt</th>
<th>Information Process</th>
<th>Recommender System Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Birukov Alexander et al, 2005)</td>
<td>Learner past behaviors and the keywords he/she has entered.</td>
<td>Factors included in the system are inter-related. Their knowledge are shared in order to enhance quality of recommendations using similarities in learners’ behavior</td>
<td>Learning Content</td>
</tr>
<tr>
<td>(Hsieh Tung-Cheng, 2010)</td>
<td>A learner with aid of other learner is entered to learn a topic which has interested in. if it is a new topic which has not been discussed yet. Learner mediator, dispatches retrieval agents for gathering learning contents from internet based on desired learning topics and store them in the warehouse of contents.</td>
<td>Major components of performing procedures are: user interface, candidate course production, preprocessing of learning topic content, production of weighted and cohesive learning topic and learning topic recommendation module.</td>
<td>The best learning contents for each course unit of a learning path.</td>
</tr>
<tr>
<td>(Tai David et al, 2007)</td>
<td>Gathering data which can varies in the form of explicit data (age, gender, field) or sequence data (learner score, course score)</td>
<td>Association mining rules for extracting recommended rules of each cluster of learners.</td>
<td>Learning Content</td>
</tr>
<tr>
<td>(Liu Feng-jung, 2007)</td>
<td>Use of precious learners’ behaviors</td>
<td>Learners’ behavior mining</td>
<td>After processing received information in system, it filters document links and shows them to users. Users can click on these content links to study according to their desires.</td>
</tr>
<tr>
<td>(Liu Feng-jung, 2008)</td>
<td>Use of inquiries which are entered by learners to system. So that, at least 10 keywords, which are used in previous activities of users, are recorded. Also, their frequencies are considered.</td>
<td>Learners’ behavior mining and utilization of semantic web technology.</td>
<td>The most adaptable contents to users</td>
</tr>
</tbody>
</table>
Basic Concepts
In this section, we are being familiar with definitions, terms and basic concepts which are related topics accordance with areas of the proposed architecture. Being familiar with these definitions make them more understandable in following sections.

Clustering
Clustering techniques can be counted as an effective tool for clustering students according to their similarities. K-Means as a clustering method is simple; nevertheless, it is a basic way to analyze other methods including fuzzy clustering.

The simplest way to implement K-Means Algorithm is firstly to select members to the number of clusters required, randomly as centroids. Then, other members based their similarity measures to these centroids are lied in distinct clusters and thus the first clusters are formed. Now, in each produced cluster using average distance function, central node (member) is calculated and repeatedly distances of member from new centroid are processed using K-Means algorithm. In this step, some of the member may transmit to other clusters. This procedure continues until there are no changes in distances of members. Formula 1 is a target function to minimize distance called Euclidean function.

\[
[d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}] \tag{1}
\]

In this Formula, the vectors \( \mathbf{x} = (x_1, \ldots, x_n) \) and \( \mathbf{y} = (y_1, \ldots, y_n) \) includes attributes of two vectors. One of these vectors is related to attributes of the central member and the other is relevant to a member to which distance from the central node is calculated.

Association Rules Mining
Association Rules represents relations between items based on their occurrence pattern in transactions (unsorted). In web transactional context, association rule are used to show the relations of web-page visits based on learner patterns.

The goal of mining using association rule is to find relations among components of a set. In this case, searching and finding interdependencies, solidarities and existing cause and effect structures among a series of components or objects in relational database and warehouse take place in this category. This rule is as following:

Prior \( \rightarrow \) Poster [Support Coefficient, Confidence Coefficient]

Support Coefficient is a probability of prior and poster inclusion in a transaction.

Confidence Coefficient is a probability that if a transaction meets conditions of prior, then also meets conditions of poster. This coefficient is represented by C sign.

Recommender Systems
With considerable growth of internet and expansion of information within that, the problem of information redundancy is being appeared more. Hence, information available on web is more than capability of users in management, absorption and maintenance of them. One of the solutions to this issue is personalized or recommended systems. These systems attempt to offer users’ favorable items.
The most popular recommender systems used and produced nowadays are categorized as Collaborative Filtering. In this method, at first, information regarding to users are cumulated and after identifying the similarities between users, they will be notified about recommendations.

**Proposed Architecture**

In this section, we analyze the architecture of proposed system and its major components

**AELTRRec Architecture**

Proposed architecture is explained in figure 1. As you can see, this architecture has been composed of three major parts. They are AELTRRec User Interface and Recommender System. In the user interface, each of learners can interact with the system using this module.

Recommender: it includes a recommender system for personalizing learning contents that interconnect with three databases. These databases are Content Rate, learner profile and learning contents. The procedure of recommender system is to input learners’ profiles and the marks learners gave to each of the contents; based on result obtained by this process, suggest learners appropriate learning contents.

![AELTRRec Proposed Architecture](image)

**Learner Modeling**

One of the important sections in implementation of the architecture is learner modeling in this system. One of the existing models for learner modeling is PAPI (Public And Private Information). Descriptions of this model by IEEE LTSC as a data transfer protocol have been designed for communication between collaborative systems. It describes learner’s information (OunnasAsma, 2005).

PAPI standard reflects Intelligent Tutoring Systems in which information related to learner’s performance is described as vital information about him/her (PanevaDesislava, 2004).

This model allows viewing learners’ information in different perspectives. (Known perspectives can be learner, instructor, parents, school, employee and others) to some extent privacy and security are considered. One of the vital clues of this standard is logical division which separates security and management of different kinds of learner information. This standard is depicted in Figure 2 which is categorized in six parts covering different parts (Jerman-BlazicBorka and KlobucarTomaz, 2005).
In the proposed architecture, for extracting input parameters and completion of learner’s profile, PAPI learner model will be used.

![High-Level Architecture of PAPI Profile](image)

**Figure 35. High-Level Architecture of PAPI Profile**

**Description and Preprocessing of Data**
Each of input attributes in the proposed architecture has different importance in clustering process. Also, each attribute value has variant domain. Hence, extracted row data related to each learner must be processed before clustering. As it is mentioned, domain value of each attribute is varied; for example age of people in data gathered for testing recommended algorithm might be ranged from 22 to 48. But, gender of attendance can be 0 (male) and 1 (female). So it requires that values of all attributes are converted to a domain of values. New values for each attribute are calculated according to Formula 2:

\[ v_\text{a}' = \frac{v_\text{a} - v_\text{amin}}{v_\text{amax} - v_\text{amin}} \]

In this formula \( v_\text{a}' \) is the value of attribute \( \text{a} \) and \( v_\text{a} \) is the converted value of attribute \( \text{a} \). \( v_\text{amin} \) and \( v_\text{amax} \) are minimum and maximum values of attribute \( \text{a} \). For example, if age of people in data gathered ranged from 22 to 48 and we are willing to normalize 35 years of old, according to above formula we have:

\[ v_{35} = \frac{35 - 22}{48 - 22} = 0.5 \]

**Learner Clustering**
At the beginning, in the recommended system with the intention of suggesting contents to learner, it needs to be clustered. In this solution, K-Means as clustering algorithm will be used. In this algorithm several factors are considered which exist in learner’s profile that is accordance with PAPI model and learner’s interested criteria for self-regulated learning. In order to cluster and help learners to become self-regulated learners, criteria must be stipulated which are in the learner’s profile and are asked him/her before course is started.

After all the values are normalized, there must be a weight associated with each of attributes which describes its degree of importance. This work is done so as to specify effects of different attributes on learners clustering.

Based on weight concept, vector \( \overline{V}_\text{a} \) indicating learner’s attribute is presented as formula 3:

\[ \overline{V}_\text{a} = (v_{\text{a1}}, w_1), (v_{\text{a2}}, w_2), \ldots, (v_{\text{am}}, w_n) \]

In vector \( \overline{V}_\text{a} \), \( v_{\text{ai}} \) to \( v_{\text{am}} \) represent attributes of learner \( \text{a} \) and values of \( w_1 \) to \( w_n \) indicates weights of learner’s attributes. This vector for learner \( \text{b} \) is presents as formula 4:

\[ \overline{V}_\text{b} = (v_{\text{b1}}, w_1), (v_{\text{b2}}, w_2), \ldots, (v_{\text{bn}}, w_n) \]
Difference between learner a and b is calculated using formula 5 (Jin Du et al, 2006):

$$d(a, b) = \sqrt{w_1^2(v_{a1} - v_{b1})^2 + w_2^2(v_{a2} - v_{b2})^2 + \cdots + w_n^2(v_{an} - v_{bn})^2}$$

As it is obvious in formula 5, importance level of a feature in learners’ cluster boosts with increase in value of the weight associated with.

Results obtained from applying clustering algorithm on a dataset according to selections of algorithm’s parameters, are so different. The goal of validating clusters is to find clusters which mostly relate to their data.

In this part, silhouette coefficient is introduced which allows selecting the best value of K in K-Means algorithm:

The idea behind this coefficient is that components within a cluster are most similar to their representative; average distance between O and components with the cluster is calculated using Formula 6:

$$a(o) = \frac{1}{|C_o|} \sum_{p \in C(o)} d(o, p)$$

In this formula C represents a cluster including o as one of its members. P shows other members in the cluster. |C_o| shows the number of members in the cluster and d represents distance function.

Components in different clusters must be dissimilar to each others. Average distance between o and the second closest cluster are calculated using Formula 7:

$$b(o) = \min_{C_i \neq C_o} \left( \frac{1}{|C_i|} \sum_{p \in C_i} d(o, p) \right)$$

According to Formula 7 and 8, the aforesaid coefficient for o is calculated by Formula 8:

$$s(o) = \begin{cases} 1 & \text{if } a(o) = 0, \text{i.e. } |C_o| = 1 \\ \frac{b(o) - a(o)}{\max(b(o) - a(o))} & \text{else} \end{cases}$$

Value of this coefficient ranges between +1 and -1 and if the value is more close to +1, it will indicate that assigning o to its cluster is an excellent assignment.

It is worth mentioning that, in developing the Silhouette algorithm, value of coefficient for k=2 to k equals to a half of the number of input data calculated and the value of k which is selected so that maximum possible value of this coefficient is produced, is selected for clustering.

**Extracting Association Rules and Recommend Contents to Learners**

After clustering learners, using Apriori algorithm and based on scores learners gave to the contents read before in the system, association rules regarding to each cluster of learners must be extracted. According to extracted rules for each cluster of learners, system recommends contents to learners in the cluster.

**How to Test and Extract Results**

For extracting results in this work, first we should select among learners to work with the developed system based on the proposed architecture. Also an electronic course must be opted for teaching the learners. It is assumed that for learning content selection, resources available on web and internet are used dynamically. These learners will work with the system for ten days to extract desire data such as similar learners, contents viewed by each learner, scores given by each learner to studied contents and other similar items for recommending content to other learners. After this step, derived from the proposed architecture and described procedures, right contents to right learners are recommended.

Before studying contents, one pretest embedded in the system is to be done which relates to each of the educational goals. After finishing course, one posttest is hold. It seems that the
proposed architecture is good enough to enhance learner’s performance; it is expected that average progression percentage of learners from pretest to posttest, in the case that contents are suggested to learner over run the case where it is free to select contents.

Conclusion and Future works
As it is discussed in this paper, recommender systems in e-learning have significant effects on improving learners’ functionalities. In this paper, with analyzing of recommender systems in e-learning, using one of clustering algorithm and association rules techniques, a recommender system was proposed.

Research area related to subject of this paper includes:
- One of the useful topics in context of recommender system in e-learning is assessment of its effect on self-regulated learning. Combining self-regulated theory with those items which should be considered in recommender systems, leads to presenting a useful recommender system which in addition of suggesting suitable contents exclusively for each learner, also enhancing its abilities in self-regulated learning (Chen Chih-Ming, 2009).
- Over and above, suggesting contents to learners, it is possible to providing them with guidance during their learning sessions. Work on adaptive and personalized guidance is a topic of great research in future.
- Another context which being raised by this paper, is considering a new concept titled “attack on recommender systems”. The purpose of this attack is misuse of open nature of recommender systems and entering unreal information (ex. scoring items) with aim of altering functionalities of recommender systems (Bhaumik, R. et al, 2007).

References
Birukov Alexander, Blanzieri Enrico, Giorgini Paolo: Implicit: A recommender system that uses implicit knowledge to produce suggestions, University of Trento, pp.1-8, April 2005.