Adaptive algorithm based on clustering techniques for custom reading plans

MARYLIN GIUGNI, FRANCISCA GRIMÓN, LUIS LEÓN
(Faculty of Science and Technology, University of Carabobo, Valencia, VENEZUELA
{mgiugni, fgrimon, lleon}@uc.edu.ve)

JOAQUÍN FERNÁNDEZ, JOSEPH MONGUET
(Multimedia Applications Laboratory, Polytechnic University of Catalunya, SPAIN
{jfernandez, jm.monguet}@upc.edu)

Abstract: Individuals use Information and Communication Technologies (ICT) to relate remotely to each other, perform any sort of transactions, and produce and assimilate large volumes of information, among other things. This has led information repositories in digital format to grow exponentially. At the same time, accessing large volumes of information and selecting the closest one to the user’s interests is increasingly difficult. With the aim of facing this problem, a tool oriented toward the personalization of readings plans in a learning environment, was developed with a view to assessing its effectiveness and the user’s satisfaction vis-à-vis the proposed adaptation algorithm. This application is based on information recovery techniques, clustering algorithms and adaptation rules to adjust the user’s model to contents and objects of study. The initial results reflect the effectiveness of the system and the users’ acceptance degree.

Keywords: Information recovery, Reading plan, Adaptation algorithm, Learning environment, Clustering algorithms, Personalization.

1 Introduction

Adaptive hypermedia is applied to construct adaptive learning paths in learning management systems based on learners’ skills, needs, and behavior [1][2].

On the other hand, there is an increasing interest in applying data mining to educational systems [3]. Data mining techniques can help to extract or detect hidden user characteristics and behaviors from large databases.

Clustering is a data mining technique, also called segmentation or unsupervised learning, useful for discovering interesting data distributions and patterns in the underlying data. It is a process of grouping physical or abstract objects into classes of similar objects [4][5].

The essential motivation for using these techniques is that they have demonstrated to work really well for modeling preferences of users [6].

The purpose of this research was to develop a tool oriented toward the personalization of readings plans in a learning environment, according to [7], which provides the user desired information, minimizing the search effort and the amount of irrelevant information. This tool, call A2plans categorizes information from the Web, restricts the search space and adapts the content; modifies the user profile, according to their interaction with the system in order to increase the relevancy of the information provided by A2plans.

The paper is organized as follows. Section 2 depicts the research methodology used. We present the system architecture in Section 3. In Section 4, experiments on the real-world data sets are conducted to evaluate the recommendation. Section 5 concludes the paper.

2 Methodology

We used the empirical research methodology called "Action Research"[8], and a combination of complementary quantitative and qualitative methodologies, which help to offset the biases of the study methodology. Action research involves a cyclical process resembling a spiral of steps; each circle consists of identification of the needs, formulating a plan of action (planning), implementing the plan (action), and fact-finding about the results of the plan (evaluation) in order to concurrently solve problems and create new knowledge. Based on the evaluation, the plan is revised and a new plan is implemented, and hence another cycle begins.

On the other hand, we used the reference model CRoss-Industry Standard Process for Data Mining (CRISP-DM) [9][10], to cover the software development phase, in the field of data mining.

2.1 Participants

The sample was made up by 26 undergraduate students who were in their final year of study, and 6 teacher-researchers; the all live in Venezuela.
2.2 Identification of the needs

The initial phase of the research focused on understanding the problem. During this stage there was a literature review of the research areas: personalization, recommendation, data mining and adaptive educational systems. Furthermore, skilled questionnaires and interviews were conducted with the purpose of knowing the problem.

During the literature review some models were examined in Information Retrieval (IR). Various retrieval models have been developed and investigated over the past decades based on a variety of mathematical frameworks [11]. The majority of IR systems are based on the Vector Space Model (VSM) [12]. This model, the most frequently used one in experimental environments, [13] represents documents and queries as vectors in a multidimensional space.

Brusilovsky says that “the web has been an important media to develop education experiences and more of them use hypermedia adaptive techniques to personalize the learning process” (cited by Brown) [14]. The aim of the hypermedia adaptive systems is to build a learning space that can adjust to each student’s characteristics, to configure educative environments in which the students reach the learning objectives through contents and paths adapted to their aptitudes, interests and preferences [15].

Adaptive systems provide a general model to represent user’s profile, an important factor when reading customize plans. The techniques and models used in IR and AHS provide bases for the construction of knowledge managed by A2plans.

On the other hand, the questionnaires answered by 26 undergraduate students showed that 86.66% of them spend 16 to 22 hours a week in search of digital information, of which 20% (5 hours/week), are used for classification, 92% said the information retrieved meets between 25% and 35% of their requirements; 100% of respondents said that they should perform searches with different criteria, to obtain results according to their interests.

The 6 interviews with teacher-researchers reflected the need to provide a tool to personalize the search results tailored to each user profile.

In summary, the diagnostic evidence of the importance of providing applications to automate the retrieval of information and provide documents relevant to their interests.

2.3 Planning and implementation of the action

The results of the analysis showed the importance of reducing response times in the classification of the recovered data, and provide custom reading plans to help reduce information overload and the student's cognitive load. Considering these needs was designed and developed A2plans.

3 A2plans Architecture

A2plans’s architecture (Figure 1) is based on a Contents Model, a User Model and an Adaptation Model. It uses an adaptation algorithm [7] to adjust the user model in accordance with the users’ evolution and to provide documents adapted to their interests.

As far as contents are concerned, the system groups documents using the K-means clustering algorithm [16]. This way, clusters consisting of documents that are similar to each other are obtained. This algorithm is used to create clusters of users and documents. K-means has been successfully used in other investigations [17] [18].

The adaptation algorithm designed for this system considers different aspects including: the student’s performance, knowledge level, document location in each cluster, feedback user, among others. This algorithm adjusts the user model according to the evolution of the learner and retrieves relevant information related to his/her profile and knowledge.

According to [19] and referenced by Ochi et al., [20] recommendation systems have three main components: background data, which is already in the system before the commencement of the recommendation process; input data, which the user gives the system in order to elicit a recommendation, and an algorithm that combines background and input data to generate recommendations. Fig.1 shows the three components mentioned above and the basic functionality of A2plans.

The system developed initially stores students’ static information, which is represented in their personal data and research areas. Then, the application identifies the user’s information needs relative to the topics provided by the system; this information is used to place the student according to the initial knowledge level, which
can be beginner, intermediate, advanced or expert. As students advance throughout their education process, statistical data is generated concerning interactions and knowledge level acquired.

The recommendation provided by the system, i.e. reading plans are generated from 5 phases: 1) representation of the documents, 2) representation of user interests, 3) grouping of documents and users, 4) creation of reading plans, and 5) user feedback.

1) Logical representation of documents: the documents are initially preprocessed for further analysis. At this stage of cleaning, are eliminated accents, images and other special characters, then the documents are represented numerically using the VSM. Vector-based IR methods represent both documents and queries with high-dimensional vectors, computing their similarity by the Euclidean distance [16]. The components of those vectors are term weights; here each document is represented by a weight vector.

The scheme Term Frequency Inverse Document Frequency (TF-IDF) [12][21] is used to calculate the weight of each keyword, which determines how relevant a term within a document. A document \( d_i \) is represented by a set of words \( (t_1, t_2, ..., t_n) \), where each \( t_i \) corresponds to a keyword that describes the document and \( n \) is the size of the set; so each word \( t_i \) has an associated \( w_{ij} \). The weight of the word \( t_i \) in document \( d_j \), is shown in Equation 1 [12][22]:

\[
w_{ij} = f_{ij} \times \log(n / df_i)
\]

In Equation 1, the term \( f_{ij} \) represents the number of occurrences of term \( i \) in document \( j \); the total number of documents is given by \( n \); \( df_i \) is the number of documents in which the term appears \( i \); and \( w_{ij} \) is the weight of term \( i \) in document \( j \).

2) Representing the interests of the user: Is related to the user profile, which describes the student's interests and goals. This profile forms initially as user interest in a topic of research, interaction with the system and the results of the diagnostic evaluation. These parameters can be configured according to the domain where the system is used.

The initial profile is used to generate a reading plan according to user needs, thus the system can retrieve and display documents that make up that recommendation. The user profile will be updated as to qualify each of the documents, according to their contribution to their research area and for this we used a Likert scale level 5.

3) Grouping of documents and users: Once the documents have been characterized, A2plans groups the documents using the clustering algorithm k-means [16], thus obtained groups made up of documents that are similar. A2plans is an easy and simple to divide the documents into k groups of similar documents in the same area of research, where \( k \) represents the number of research areas. The main idea is to define \( k \) centers (one for each group); these \( k \) centers represent the most relevant documents for each area of research.

However, once it has established a number of groups, \( k \) documents are selected to be the initial centers of each cluster. This step is important to select documents whose content has a high correspondence with the research area, therefore if the documents are not appropriate, the classification could be inefficient.

Each time a new document is added to the system, is located in the closest group to this Euclidean distance is used, i.e. the distance between its feature vector and each of the \( k \) centers (Equation 2).

\[
dis = \sqrt{\sum_{i=1}^{n} (c_{ij} - D_j)^2}
\]

In Equation 2, \( c_{ij} \) corresponds to the characteristic vector of the center \( k \). \( D_j \) is the characteristic vector for document \( j \), \( n \) is the size of vector, and \( dis \) represents the Euclidean distance [16]. When a document is added to one of these groups, it recalculates the center of all clusters and the distance of all other documents concerning them, to see if any of the old records should be reassigned.

4) Generation of reading plans: Once the clusters have been formed, the system generates the pl, reading plan represented by a set of documents \( (d_1, d_2, ..., d_n) \), adapted to user interests \( j \), where \( w \) represents the number of documents that make up the reading.

To determine the reading plan were considered in 3 aspects: the similarity between the profile of the user and system documents, user interaction with the system, and their interest in various areas of knowledge. We used the adaptive function (Equation 3), designed by the authors:

\[
f(\text{adaptation}) = \sin(T - 3)^2 + \ln(R) \times 4
\]

In equation 3, \( T \) is for the interest of the user in a paper, \( R \) refers to the user interaction. The sine function, provides a positive and a negative part to the possible value of the documents; is responsible for reflecting the interest or disinterest in the areas administered by the system. In addition, the natural logarithm considers the participation in the forums and collaborative work packages provided by the system.

5) User feedback: This phase corresponds to the explicit feedback that the user performs \( f \) to each of the documents that make up your reading plan. The values of feedback in the system are between zero and five inclusive.

As the user evaluates the documents, modifies the user profile, and generates a reading plan according to your preferences. The profile is updated using Rocchio's...
formula (Equation 4) [23] [24], which is based on the relevance of the documents reviewed by the user. To A2plans a document is relevant if its valuation is greater than or equal to 2 points (on a scale from zero to five).

$$C_i = \alpha C_0 + \beta \sum_{j=1}^{n_r} R_{ij} - \gamma \sum_{j=1}^{n_{ir}} \frac{NR_{ij}}{n_{ir}}$$ (4)

In Equation 4, C0 represents the characteristic vector of the initial profile of the user. Ci represents the characteristic vector of the user profile after an iteration. Rij is the characteristic vector of the relevant document i. NRij is the characteristic vector of the relevant document i, nij represents the number of relevant documents, nir the number of irrelevant documents. The constants α, β and γ to adjust the impact of relevant documents and irrelevant, these values were adjusted following the suggestion by [25], so α has the value 1, β of 0.75 and γ of 0.25.

4 Description of the experience

A2plans was evaluated by 26 undergraduate students. The experiment began with 300 documents type Portable Document Format (PDF), all framed within the four areas of knowledge of computer science.

A2plans shows the user the documents ordered by the level of significance, showing the documents whose ratings are higher and that best fit your search profile, the latter is achieved by determining the proximity of the user profile for each document. Importantly, reading plans are updated as the documents are evaluated by various researchers. Positive or negative evaluations, affect the relevance of the documents in a given research area.

To measure the effectiveness of the personalization system, it was subjected to a qualitative assessment by calculating the reading plan precision. Moreover, to supplement these preliminary results, questionnaires that allow for knowing the user’s satisfaction with respect to the different parts of the system have been employed. These questionnaires are another alternative to measure the application quality.

With regard to user satisfaction in relation to using the system and in particular the recommendation of reading plans, a questionnaire [26] [27] was used by 12 users that make up the sample and the observed data in Table 1.

In relation to the results, it appears that by adding the data in columns “high degree of agreement” and "agreement", the items 1, 5, 8 and 9, 81.73% of the sample considers that generates A2plans reading plans tailored to user interests. Although on average 16.15% of respondents felt indifferent results generated by the system, the 83.46% were "agreement" and "high degree of agreement” in the results of the system. Also, 100% of users said the system is easy to use.

<table>
<thead>
<tr>
<th>Aspects evaluated</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content adjusted to the profile</td>
<td>0.00</td>
<td>0.00</td>
<td>30.77</td>
<td>42.31</td>
<td>26.92</td>
</tr>
<tr>
<td>Usefulness of content</td>
<td>0.00</td>
<td>0.00</td>
<td>38.46</td>
<td>38.46</td>
<td>23.08</td>
</tr>
<tr>
<td>The categories generated are appropriate to the needs of information</td>
<td>0.00</td>
<td>7.69</td>
<td>19.23</td>
<td>30.77</td>
<td>42.31</td>
</tr>
<tr>
<td>Contribute to solving the information needs</td>
<td>0.00</td>
<td>0.00</td>
<td>15.38</td>
<td>30.77</td>
<td>53.85</td>
</tr>
<tr>
<td>Reading plans tailored to the needs</td>
<td>0.00</td>
<td>0.00</td>
<td>19.23</td>
<td>34.62</td>
<td>46.15</td>
</tr>
<tr>
<td>Ease of use</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>3.85</td>
<td>96.15</td>
</tr>
<tr>
<td>Relevance of documents recommended by the reading plan</td>
<td>0.00</td>
<td>0.00</td>
<td>11.54</td>
<td>26.92</td>
<td>61.54</td>
</tr>
<tr>
<td>Plans evolve according to the user profile</td>
<td>0.00</td>
<td>0.00</td>
<td>23.08</td>
<td>46.15</td>
<td>30.77</td>
</tr>
<tr>
<td>The content of the retrieved documents relate to the domain of study</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>19.23</td>
<td>80.77</td>
</tr>
<tr>
<td>Recovered documents are highly relevant to the category you are</td>
<td>0.00</td>
<td>0.00</td>
<td>15.38</td>
<td>61.54</td>
<td>23.08</td>
</tr>
</tbody>
</table>

Table 1: Results of the questionnaire (data expressed in percentage), legend: (1) total disagreement, (2) disagree, (3) indifferent, (4) agreement, (5) high degree of agreement

On the other hand, considering the importance role of the user in such systems, interviews with them, they were crucial to identify strengths and weaknesses of the system and that they showed the importance of taxonomy keywords. In addition, there were criteria to evaluate each user as relevant or not relevant document.

To evaluate quantitatively the quality of the clustering of the documents within the cluster average similarity was used [28] [29] [see Equation 5], which is a measure that determines the level of similarity between the documents within the cluster depending on their characteristics.

$$SP_j = \frac{1}{n_j} \left( \sum_{i=1}^{n_j} \sum_{j=1}^{n_j} \text{sim}(d_i, d_j) \right)$$ (5)
This value ranges between zero and one, where the similarity is close to one, meaning that documents are strongly associated, i.e. the cluster is very homogeneous, while values closer to 0 express a gradual dissimilarity.

Once the evaluation of the 4 cluster, is shown in Table 2 that the average similarity between cluster 2 is very close to one and in general this measure exceeds 0.75 in each cluster, indicating that the documents each group are very similar.

<table>
<thead>
<tr>
<th>SP₁</th>
<th>SP₂</th>
<th>SP₃</th>
<th>SP₄</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.754</td>
<td>0.894</td>
<td>0.817</td>
<td>0.761</td>
</tr>
</tbody>
</table>

Table 2: Average similarity of the 4 cluster (equation 5)

5 Conclusion

The purpose of this paper was to propose a recommendation system documents adapted to user interests, in which clustering algorithms are used to determine the similarity in the behavior of users and provide a collaborative search space. Furthermore, this technique was used to sort and partition the documents by reducing the search space, and thus facilitate the administration, comprehension and understanding of content provided by A2plans.

Based on the experiments conducted, most of the students will accept the reading recommendations the system offers. The recommendation system reads proposed has proved to be both workable and effective.

The experimental study has found that reading plans generated by the system are consistent with the interests of the user and have been rated positively by it.

Although quantitative assessments made to the system, have been satisfactory, it is necessary to make a further assessment, to validate these results.

The best results could be related, in part, to the sensible decision to include the retrieval of documents using clustering techniques. This suggests that it is important to properly select the elements contained in the adaptation function.

In the educational context, this system not only helps students recommending documents tailored to their needs and by working groups, but also helps the teacher by providing a tool to monitor student progress, which may suggest reading plans remediation if necessary.

The integration of data mining techniques with AHS was effective, which has led to a system that can be used in different research domains. The result of this research work benefits the community of application developers in the area of information recovery in adaptive hypermedia systems, who desire to put this experience into practice.

The authors consider this paper as an initial model for integrating data mining techniques for content adaptation in learning environments. A future study should focus on designing and developing of MultiAgent Systems (MAS) that allows coordinate the workflow of a learning environment.

References: