Int. J. Human-Computer Studies 76 (2015) 22–39

Contents lists available at ScienceDirect

Int. J. Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Evaluation and selection of group recommendation strategies for collaborative searching of learning objects ☆

Alfredo Zapata [a], Víctor H. Menéndez [b,*], Manuel E. Prieto [c], Cristóbal Romero [d]

[a] Autonomous University of Yucatan, Faculty of Education, 97150 Mérida, Mexico
[b] Autonomous University of Yucatan, Faculty of Mathematics, 13615 Mérida, Mexico
[c] University of Castilla-La Mancha, Computer Science Faculty, 13071 Ciudad Real, Spain
[d] University of Cordoba, Dept. of Computer Science, 14071 Córdoba, Spain

ARTICLE INFO

Article history:
Received 17 July 2013
Received in revised form
26 November 2014
Accepted 9 December 2014
Communicated by E. Motta
Available online 18 December 2014

Keywords:
Learning objects
Group recommendation systems
Collaborative web search
Metadata
Repository
Meta-learning

ABSTRACT

Nowadays, there is a wide variety of e-learning repositories that provide digital resources for education in the form of learning objects. Some of these systems provide recommender systems in order to help users in the search for and selection of the learning objects most appropriate to their individual needs. The search for and recommendation of learning objects are usually viewed as a solitary and individual task. However, a collaborative search can be more effective than an individual search in some situations – for example, when developing a digital course between a group of instructors. The problem of recommending learning objects to a group of users or instructors is much more difficult than the traditional problem of recommending to only one individual. To resolve this problem, this paper proposes a collaborative methodology for searching, selecting, rating and recommending learning objects. Additionally, voting aggregation strategies and meta-learning techniques are used in order to automatically obtain the final ratings without having to reach a consensus between all the instructors. A functional model has been implemented within the DELPHOS hybrid recommender system. Finally, various experiments have been carried out using 50 different groups in order to validate the proposed learning object group recommendation approach.

1. Introduction

Web searching is generally considered to be a solitary activity, and all major search engines and Web browsers are designed for solo use. However, many tasks in both professional and casual settings can benefit from the ability to jointly search the Web with others (Morris, 2008). In the educational domain, for example, a group of instructors in the same course may be interested in searching for and selecting together the educational resources most appropriate to develop a new digital course. That is, instead of only one instructor being in charge of developing the courseware, or several of them developing different parts of the courseware, all the instructors in the course would be together searching and selecting the resources to be used in the course. This type of searching in a group is called a "collaborative search" and it allows many individuals to benefit from preferences and experiences of other like-minded individuals (Smyth et al., 2011).

A learning objects (LOs) search is one of the most time-consuming tasks for instructors, because finding the most appropriate objects to match a specific subject is not always easy. An LO is a basic component (unit of a course) or modular digital resource that can be used to support learning (Wiley, 2002). For example, Fig. 1 shows the interface of an LO on learning the English language, which is oriented to learners between 12 and 20 years old in order to promote their reading and comprehension skills.

A vast amount of LOs are published and distributed across the Internet in repositories such as MERLOT (Schell and Burns, 2002), MACE (Stefaner et al., 2007), AGORA (Prieto et al., 2008), ARIADNE (Ternier et al., 2009), etc. Facing huge volumes of LOs, instructors may be lost when selecting the most suitable LOs to be used in a specific course. In order to resolve this problem, recommender systems have been successfully applied to provide suggestions about LOs which can be most useful to the individual knowledge, goals and/or preferences of each user (Manouselis et al., 2011). However, there are situations when it would be good if we could recommend to a group of users rather than to an individual. The main problem in group recommendation is determining how a group of people reaches a consensus about the score for each item in a way that reflects the interests and preferences of all group members. Nowadays, there are some examples of group

☆This paper has been recommended for acceptance by E. Motta.

* Corresponding author. Tel./Fax: +52 9999 423140.
E-mail addresses: zgonzal@uady.mx (A. Zapata), mdoming@uady.mx (V.H. Menéndez), Manuel.Prieto@uclm.es (M.E. Prieto), cromero@uco.es (C. Romero).

http://dx.doi.org/10.1016/j.ijhcs.2014.12.002
1071-5819/© 2015 Published by Elsevier Ltd.

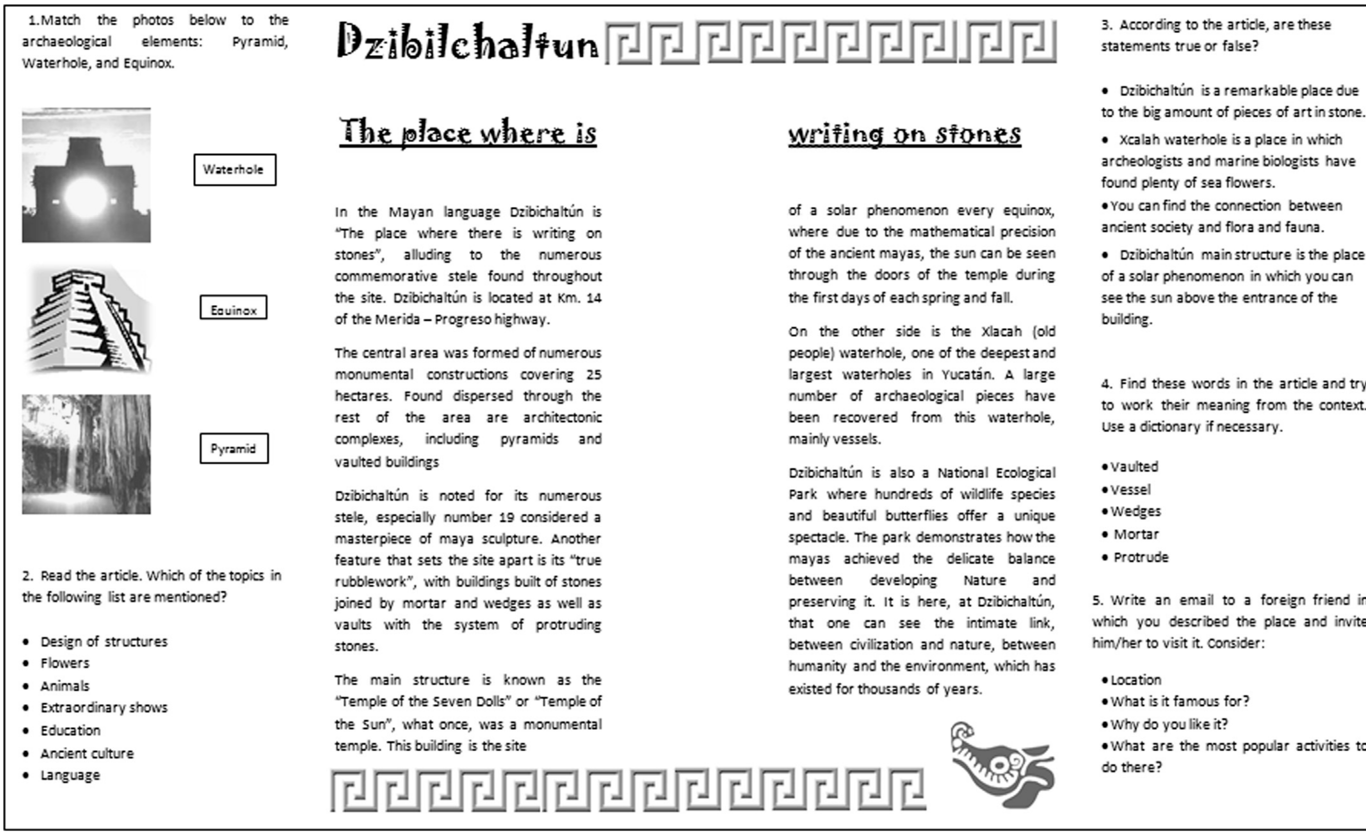

Fig. 1. Example of learning object on learning the English language.

recommenders for selecting television programmes for a group to view or for selecting a sequence of songs to listen to in a group (Masthoff, 2011). However, in the current bibliography, we have not found any group recommender systems in the education domain or specifically for LO repositories.

This paper proposes a collaborative methodology for searching, selecting and rating LOs in a group. We have implemented this methodology into a hybrid recommendation system called DELPHOS (Zapata et al., 2013), which is a framework to assist users in the single/individual personalised search for learning objects in repositories (http://smile.esi.uclm.es/delphospruebas/). We have extended DELPHOS with new functionalities, including the creation and management of groups of users, the realisation of collaborative activities, and the recommendation of the most interesting LOs to these groups. We also propose a meta-learning approach in order to help the mediator of a group to select the best rating aggregation method depending on the rating of previous similar groups. For one thing, the mediator is free to use any of the available aggregation strategies to automatically obtain the LO ratings from a group of users, without needing to use the traditional democratic in-person or online discussion to obtain a consensus from all the group members about each LO. But also, the mediator can directly use the best aggregation strategy recommended for a group based on its characteristics. In this way, the traditional time-consuming consensus-taking among users can be avoided by using an automatic method based on meta-learning and voting aggregation strategies.

The remainder of the work is organised as follows: Section 2 describes the related background and the most similar works; Section 3 describes the proposed methodology; Section 4 shows the implementation of this methodology within the DELPHOS system; Section 5 describes the experiments and results for validating the efficiency of this system with groups of real users; finally, Section 6 outlines some concluding remarks and future research lines.

2. Background and related work

This work deals with the specific problem of evaluating and selecting voting aggregation methods for assisting the search for the best LOs appropriate to the interests of a group of instructors. In this section, we introduce the subjects most relevant to this work. We start with concepts of collaborative social search, then we introduce group recommender systems, and finally we explain data mining and meta-learning concepts.

2.1. Collaborative web search

Collaborative web search (CWS) is an activity in which participants work together in a synchronous or asynchronous collaboration in order to satisfy an information need (Morris, 2013). Nowadays, CWS is generally implemented around mainstream Web use rather than for LOs. It can be divided into implicit or explicit collaboration:

Implicit search engines. These are characterised by identifying similar users, queries and links clicked automatically, and by recommending relevant queries and links to the searchers. Some examples of these systems are as follows:

- *Jumper* 2.0 (Jumper Networks Inc., 2011), which empowers users to compile and share collaborative bookmarks by crowdsourcing their knowledge, experience and insights, using knowledge tags.

- *Seeks* (Seeks Project, 2011), which allows users to regain control over the selection of results of searching, and also allows a personal profile that can be shared with other users to be built.

Explicit search engines. This collaborative modality means that users share an agreed-upon information need and they work together towards that goal. Some examples are as follows:

- *SearchTogether* (Morris and Horvitz, 2007), which is a system to enable groups of remote users to synchronously or asynchronously collaborate when searching the Web. This software employs a client/server architecture. The server acts as an intermediary for sending shared state among clients, as well as being a repository for storing SearchTogether session data in order to enable session persistence.
- *Cerchiamo* (Golovchinsky et al., 2008), which is a collaborative exploratory search system that allows teams of searchers to explore document collections synchronously. In comparison to other synchronous collaborative search systems, it allows the users to pursue the search task at their own pace without needing to synchronise each search activity with their collaborators. CWS also involves some passive filtering that occurs between the two searchers, using a specialised algorithm. Therefore, this system combines both explicit and implicit collaboration.
- *CoSearch* (Amershi and Morris, 2008), which is a system developed to improve the experience of a co-located collaborative Web search by leveraging readily available devices such as mobile phones and extra mice. It is oriented to be used by a group of people gathered around a single computer to jointly search for information online.
- *Heystaks* (Smyth et al., 2009), which is a case-based Web search system, designed to work in cooperation with mainstream search engines. Heystaks is a search utility that is seamlessly integrated with leading search engines such as Google, via a browser toolbar/plug in, in order to offer collaborative features that are missing in today's Web search engines.

All the previous collaborative web search systems are general, i.e., they are not specialised in any specific domain. Therefore, they return to the user a great variety of contents (images, videos, etc.), which, in most cases, are not educational resources. And although some engines provide advanced and customised searches, an LO collaborative search engine can provide more specific filters (such as the most downloaded LOs, the best rated, the most related to an academic profile, etc.) for retrieving only the most relevant LOs according to the user's interests. It should be noted that we have not found any systems or works specifically focused on facilitating the collaborative search of LOs. However, some of the general collaborative web search systems described above are consistent with our DELPHOS system in some aspects. For example, the *SearchTogether system* coincides with our proposal to put forward a collaborative environment aimed at a group of users who know each other and who are working together towards a common objective, in both a synchronous and an asynchronous or remote manner. Another aspect of coincidence is that both systems allow for the assessment of the resources that they consider appropriate to their interests. The *Cerchiamo proposal* coincides with our system in that it allows for real-time collaboration on search tasks amongst a group of users. The *CoSearch system* is consistent with our proposal in that both interfaces allow users to do educational resource searches from mobile devices. The *HeyStaks proposal* concurs with DELPHOS to a certain degree, as both systems have an exclusive panel of activities where all group members can visualise the selected resources. Finally, DELPHOS also has additional features such as the application of weighted filters that search information, and the incorporation of a section that performs the group management. All this makes DELPHOS an innovative proposal for the collaborative search of LOs.

2.2. Group recommendation

Group recommender systems are more suited to groups than to individuals in some specific applications, such as those used for recommending music performances, tourist attractions, holiday destinations, movies and TV programmes (Berkovsky and Freyne, 2010). User groups usually arise from a shared need that allows the users to distinguish between at least four different notions of "group" (Boratto and Carta, 2010): established group (a number of persons who explicitly choose to be a part of a group, because of shared, long-term interests), occasional group (a number of persons who do something occasionally together, like visiting a museum. Its members have a common aim in a particular moment), random group (a number of persons who share an environment in a particular moment, without explicit interests that link them), and automatically identified group (groups that are automatically detected considering the preferences of the users and/or the resources available).

The goal of group recommendation is to compute a recommendation score for each item (in our case, each LO) that reflects the interests and preferences of all group members. The problem is that all group members may not always have the same tastes, and a consensus score for each item needs to be carefully designed (Amer-yahia et al., 2009). So, to recommend to user groups is more complicated than recommending to individuals (Masthoff, 2011). The main problem that group recommendation needs to solve is how to adapt to the group as a whole, based on information about individual users' likes and dislikes. A solution is to use group decision strategies or aggregation methods that are inspired by social choice theory (see Table 1), and establish different ways of how a group of people can reach a consensus (Masthoff, 2011).

Table 1
Examples of vote aggregation methods.

Method	Description
Plurality voting	Uses "first past the post": repeatedly, the item with the most votes is chosen.
Average	Averages individual ratings.
Median	Takes the median value of individual ratings.
Approval voting	Counts the individuals with ratings for items above an approval threshold (e.g., 3).
Least misery	Takes the minimum of individual ratings.
Most pleasure	Takes the maximum of individual ratings.
Average without misery	Averages individual ratings, after excluding items with individual ratings below a certain threshold (say 5).
Fairness	Items are ranked as if individuals are choosing them in turn.

However, groups are very diverse in nature, and no single group decision strategy (see Table 1) works best for all groups. A way to address this issue is to identify the inherent characteristics of different groups and to determine their specific impacts on the group decision process (Gartrell et al., 2010). According to Gartrell, there are at least three factors that may affect a group's decision:

- *Social*. Represents the relationship level between two or more group members. Since they interact with and influence each other, the group decision is affected by the strength of the social relationships.
- *Expertise*. Represents the expertise of group members. To reach a consensus, the group decision process usually involves mild or intense discussion. In this process, each group member is able to state his or her opinion based on the experience that he or she has.
- *Dissimilarity*. Represents the disagreement between any two group members. Intuitively, the closer the preference for an item between two members, the lower their disagreement about the item.

Finally, there are several examples of applications of group recommender systems. The approach proposed by Gupta et al. (2004) presents Group–Place Identification (GPI), an algorithm for automatic identification of informal social group members and group–place associations using community mobility traces. The method developed by Yu et al. (2005) proposes an adaptive in-vehicle multimedia recommendation system serving groups of users. In Campos et al. (2008), mechanisms are proposed that govern the interactions between group members and focus on automatically recommending a ranked list of new items to a group of users. In Chen et al. (2008), a methodology is proposed to predict the possible interactions among group members and to recommend items for groups of members. In Boratto et al. (2009), a group recommendation algorithm is implemented which, based on users' preferences, detects communities of similar users and predicts group preferences. In Gartrell et al. (2010), a group recommendation method is proposed that utilises both social and content interests of group members to construct the corresponding group consensus function.

After a review of all previous initiatives, we must note the importance of the absence of proposals in the educational domain, although some of the works described below share some similar characteristics with DELPHOS. For example, both Gupta et al. (2004) and Boratto et al. (2009) proposed a clustering algorithm to identify groups of users and produce recommendations for each group. Gartrell et al. (2010) generated associative classification rules to discover interesting relations or patterns between variables in group members. However, DELPHOS also applies a classification algorithm that provides a better general performance for predicting the aggregation method most appropriate for each new group. Still, Yu et al. (2005) concur with DELPHOS, because both systems have a recommendation system based on profile similarity. However, our proposal additionally implements three other filters based on content similarity, usage and quality of the evaluation. The method developed by Campos et al. (2008) implements Bayesian networks to represent how different individuals in a group interact in order to make a final choice or recommendation. In contrast, our proposal seeks to resolve the same problem through voting aggregation strategies and meta-learning techniques. Chen et al.'s (2008) programme coincides with our proposal in that both systems propose a hybrid recommender that combines collaborative and content-based methods. However, the DELPHOS system additionally implements a third method based on demographic approach. Finally, DELPHOS is the only system specifically oriented for LO recommendations to groups of users.

2.3. Meta-learning

Meta-learning helps to select or create optimal predictive models and reuse previous experience from analysis of other problems, relieving humans of most of the work and realising the goal of computer programmes that improve with experience (Brazdil et al., 2009; Jankowski et al., 2011). These methods are designed to automatise decisions required for the application of computational learning techniques. The principal goal for meta-learning is to recommend the best method for a given dataset that consists of objects described by some features. This requires the characterisation of the dataset by meta-features and the ability to estimate the similarity of new data to already analysed datasets (Duch, 2011). Given new data and a description of the goals, meta-learning systems should support decision-making in classification, regression, and association tasks, and/or provide comprehensible models of data. Therefore, meta-learning can be used to predict the best aggregation model for a new group of users based on their characteristics, and then automatically obtain a consensus about the recommendation of each LO to the group, based on the individual ratings.

Meta-learning belongs to a branch of Machine Learning (ML) that tries to replace human experts involved in the Data Mining (DM) process of creating various computational models, learning from data. When data come specifically from educational environments, the process is called Educational Data Mining (EDM), which is an emerging interdisciplinary research field for addressing important educational questions (Romero and Ventura, 2013). There are a number of popular methods within DM and EDM, and one of the oldest and most popular is classification, which is a supervised technique to predict class labels (Hämäläinen and Vinni, 2011). The idea of classification is to place an object into one class or category, based on its other characteristics. There are different approaches or types of algorithms for classification, such as

- *Rule-based algorithms*, which reveal rules; one of the most well known is RIPPER (Repeated Incremental Pruning to Produce Error Reduction), which is an optimised version of the IREP association rules algorithm, which implements rules by decision lists (Cohen, 1995).
- *Tree-based algorithms*, which reveal a decision tree; one of the most well known is the C4.5 algorithm, which bases its classifier induction from a set of decision trees, and is an extension of Quinlan's earlier ID3 algorithm (Quinlan, 1993).
- *Function-based algorithms*, which reveal a function; one of the most well known is the support vector machine, which constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space for classification (Platt, 1999).
- *Bayes-based algorithms*, which reveal a probabilistic classifier based on Bayes' theorem; one of the most well known is NaiveBayesSimple, which is a Bayes classifier algorithm with independent features (Duda et al., 2000).
- *Nearest neighbours-based algorithms*, which reveal the closest points; one of the most well known is the KNN algorithm, which proposes that a new case is going to be classified in the most frequent class that belongs to their K-nearest neighbours (Aha et al., 1991).

Finally, it is important to note that we have not found any previous work regarding the application of meta-learning in voting aggregation strategies or LO recommendation. In fact, the application of meta-learning in EDM is new and promising, but we can only cite two previous works. The first one focused on using

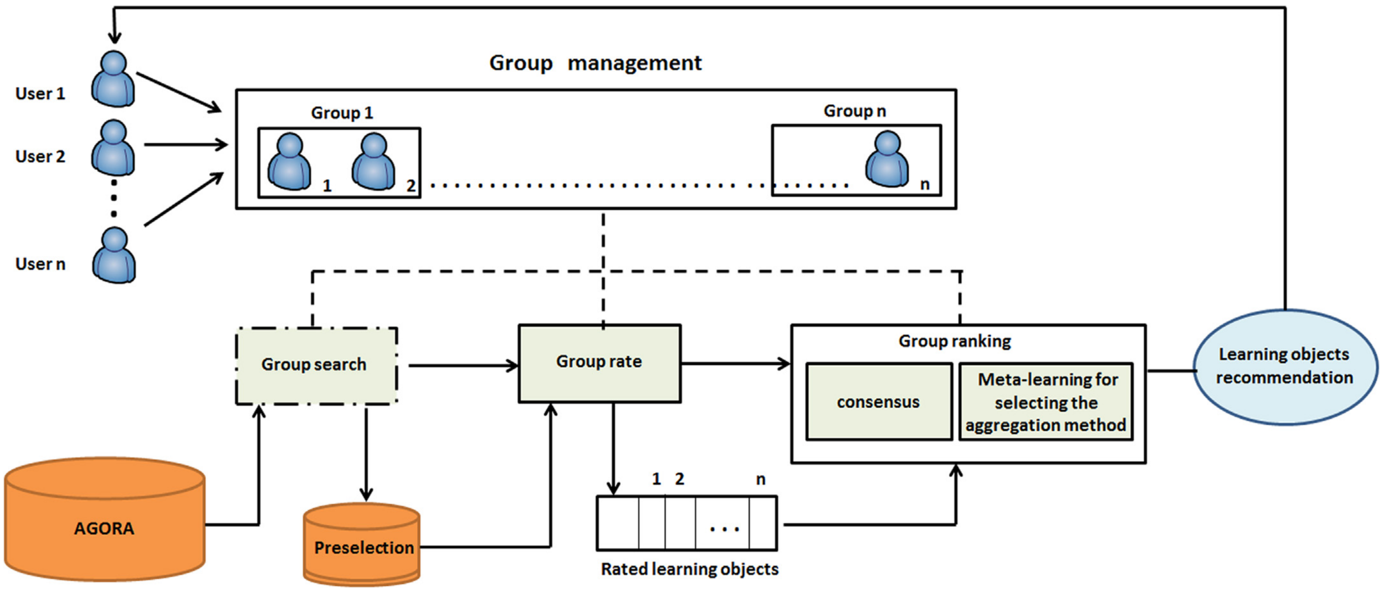

Fig. 2. The collaborative search method of learning objects.

meta-learning to support the selection of the best parameter values in a J48 classifier using several educational data sets (Molina et al., 2012), while the second one proposed the employment of several classification measures and dataset features to evaluate the performance of a set of classifiers, in order to recommend the best classifier to use with a new unseen educational data set (Romero et al., 2013). Our DELPHOS system extends this last idea by using the results obtained with the best classifier for recommending an aggregation voting strategy to the user. In this way, DELPHOS provides the mediator of the group with the possibility of automatically obtaining, through a transparent meta-learning process, the most appropriate aggregation strategy according to the group descriptors.

3. Proposed methodology

We propose a methodology for helping a group of instructors in the collaborative search for and rating of LOs on a predefined common topic or subject of interest. This methodology is based on creating groups of users and using voting aggregation strategies for recommending LOs, and it consists of four main steps (see Fig. 2): Group management, Group search, Group rating and Group ranking generation.

A necessary first step before searching for and rating LOs is to create and manage a group of users. This process starts when any user decides to create a new collaborative search group of LOs. Once created, it is possible to send requests of incorporation to other users. Next, both sub-processes are described as follows:

- *Create a group*. A user can create a group with the purpose of doing a collaborative search for LOs on a specific common topic or subject. The group creator becomes the mediator or administrator of the group, and he/she is also in charge of setting the title of the group and providing a description of the objectives or topics of interest.
- *Request of incorporation to a group*. Once a group is conformed, the mediator can send requests for incorporation to other users who have similar interests regarding the topics they search for. As example, groups can be formed among instructors of the same subject, expertise area, research group, department, etc., who are interested in searching LOs on the same specific topic. This petition for incorporation can be done explicitly by using a communication tool.

Next, each group member carries out the search step by using keywords, metadata values and different filtering criteria. They can add or recommend the discovered LO if they consider it relevant or interesting to the group according to previously defined topics. In order to facilitate these decisions and to serve as a guide to other members, each member of the group can also view all the searches done by the other group members to which he belongs, and they can add tags and comments to the added or selected LO for the group.

After all the group members have completed the preselection of LOs, they must rate or score all the LOs that have been added to the group. A typical Likert scale, where the highest value indicates the most suited LOs according to the group interests, and the lowest value corresponds to the least adequate, can be used for rating. The ratings assigned can be viewed by all group members to provide a comparison of opinions.

At the end of the ratings step, it is necessary to establish a single final group decision for all the selected LOs. There are several ways to do this. On one hand, a democracy method can be used in which all group members decide the final rating for each LO by consensus, and the best-positioned LOs (in this final ranking) will be the most relevant to the group. The problem is this method requires the communication and involvement of all group members, and sometimes it is not easy to make a decision between users with different opinions. Another way is to use a voting aggregation method, which can automatically obtain a final rating of all the LOs without needing user intervention. There is a wide range of aggregation methods described in the previous Section 2.2 (see Table 1).

However, it is difficult to determine which aggregation method is the most appropriate for each type of group. In order to resolve this problem, we propose to use a meta-learning process (Fig. 3) that obtains the aggregation method which provided/gave the best performance for a group based on its characteristics and previous rating of similar groups.

As seen in Fig. 3, the meta-learning process starts from a dataset which contains information about previous groups and their ratings. Next, the groups' characteristics are defined and the performance of the rating aggregation method is evaluated in

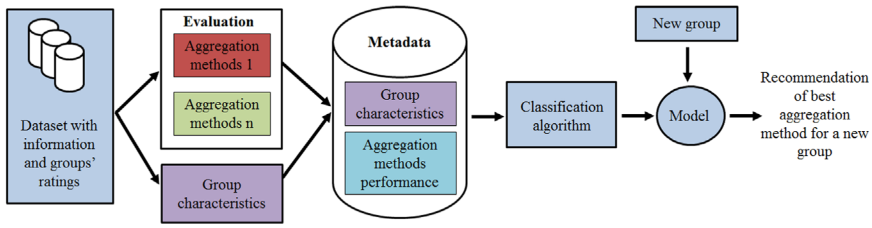

Fig. 3. Meta-learning process.

order to form a new metadata set. On one hand, based from Gartrell et al.'s (2010) work, we propose to obtain the following descriptors or characteristics:

- *Group size* – the number of members in a group. This value is automatically calculated by the system.
- *Social contact level* – the social contact level between group members. This value is obtained explicitly, as each member indicates the social relationship with each of the rest of the group members via a Likert scale. In our case, this is 0 = none; 1 = low; 2 = average; 3 = high; 4 = very high. In the case of inconsistencies among users, we use the lower value.
- *Experience level* – the experience level in teaching, and the didactic and instructional design of each group member. This value is implicitly obtained in our system, starting from the system registration user's profile of each member. In our case, this is 0 = none; 1 = initial; 2 = medium; 3 = advanced; 4 = expert.
- *Dissimilarity level* – the disagreement between the group members with respect to previous LO ratings. This value is calculated implicitly by our system from the differences between the ratings of LOs by the group members. In our case, this is a real value between 0 and 4.

Additionally, we propose a new descriptor based on the activity level of the group members in using LO repositories, because it is easy to obtain and we think that it can provide more and better information about users:

- *Activity level* – the average of interactions/activities of a group member with LOs in a repository. In our case, we have used the information provided by the AGORA repository (Menéndez et al., 2011), which is an Ibero-American learning object management system which presents a collection of services that facilitate its interoperability with DELPHOS (Zapata et al., 2013). In order to obtain the Interaction Activity value of a user ($IA_i(u)$), we used the implicit information about two type of users' interactions/activities – the publication and download frequency of LOs – by using equation

$$IA_i(u) = \frac{\sum_{i=1}^{n} (POx_i + DOx_i)}{Max(POy_i + DOy_i)} \quad (1)$$

where

- POx_i = Number of publications of learning objects (Ox_i).
- DOx_i = Number of downloads of a learning object (Ox_i).
- $Max(POy_i + DOy_i)$ = Maximum number of publications and downloads of LOs associated with a user in the repository.

Finally, *Activity level of the group members* is rescaled to a value between [0, 4].

Still, various aggregation methods can be used to automatically obtain the final group score/rating for each LO (Masthoff, 2011). In fact, we propose using the eight aggregation methods previously described in Section 2.2 (see Table 1), plus three new weighted versions of the average method based on Popescu (2013). Instead of assuming equal weights for all the members, we give more weight to some users based on their characteristics, assuming that some members are more influential and can persuade others to agree with them. These three new weighted average methods are outlined below:

- *Weighted average* (*more active user*), in which the most active users have a higher weight in the final score/rating than the less active users.
- *Weighted average* (*more social user*), in which the most social users have a higher weight in the final score/rating than the less social users.
- *Weighted average* (*more experienced user*), in which the most experienced users have a higher weight in the final score/rating than the less experienced users.

Next, an evaluation phase is necessary in order to determine which aggregation method obtains the lowest error with respect to the actual consensual final rating of group members for all LOs. This actual or real rating is the final score of the group, obtained after consensus between all the members. So, it is necessary that the group has an in-person reunion or online communication in order to achieve the final score, starting with each individual rating/score and opinion.

Next, a new metadata set is created by using the characteristics of each group and the aggregation method that provided the best group performance. Starting from this metadata set, it is possible to predict which aggregation method is most appropriate for a new group, given its characteristics. This is a classification task, where the predicted value is the aggregation method name. However, because there are a lot of classification techniques, we must therefore select a representative number of classification algorithms in order to compare their performance when using our metadata set.

Finally, the classification algorithm that provides a better general performance will be the one selected for predicting the aggregation method most appropriate for each new group. In this way, the classification model obtained by the selected algorithm will be used for selecting, in real time, the best aggregation method for a new group according to the characteristics of the group and their individual ratings.

4. Implementation

The previous proposed methodology has been implemented into DELPHOS, which is a hybrid recommendation system in LO repositories (Zapata et al., 2013). The current version of DELPHOS is associated with the AGORA repository (Menéndez et al., 2010). However, its architecture lets users incorporate other repositories in order to carry out a federation search.

In this work, we have extended DELPHOS with new features for collaborative searching for and rating of LOs. Initially, any registered user of DELPHOS can create a new group for LO searching. The user who creates the group automatically becomes

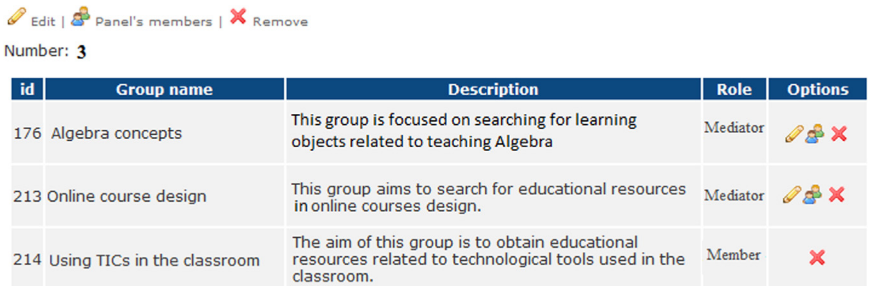

Fig. 4. List of groups to which a user belongs.

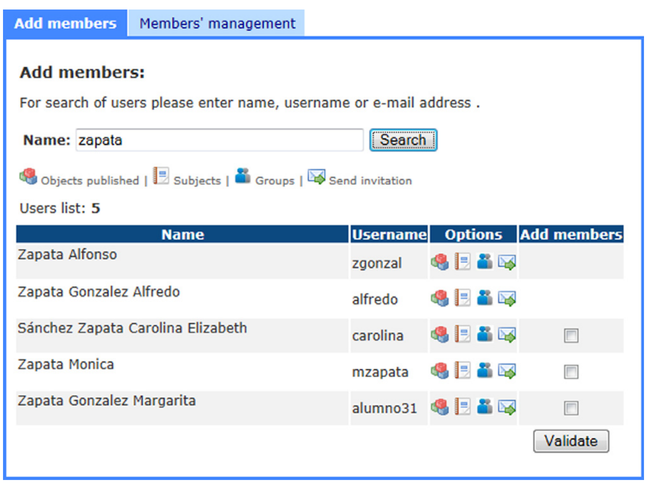

Fig. 5. Add new members to group panel.

the administrator or mediator of that group, and has to define a title or name of the group and provide a description or objectives of the group. The group mediator can also access editing functions and the panel's members (see Fig. 4).

Starting with the panel members (see Fig. 4), the mediator can perform various group management activities, including adding new members, changing members' roles, and removing members. In order to add new members to a group, the mediator must do a search of users by name, surname or e-mail. The obtained user list shows all the users that match the search parameters, together with some options (the LOs published by the users, the registered subjects, the groups to which a user belongs, and the option to send an e-mail invitation to join the group) and a checkbox for directly adding one or several users to the group (see Fig. 5). The mediator can also change the role of added members from mediator to member and vice versa, and remove members from the group.

After a group is created, the users that belong to a group can choose to do either an individual search or a collaborative social search. In both cases, the interface is the same (see Fig. 6), in which the user defines the desired search parameters based on a required text query or keywords, some optional metadata values and different filtering or recommendation criteria (Zapata et al., 2013).

Afterwards, DELPHOS shows the user a ranked list of recommended LOs (see Fig. 7) together with icons (see Table 2) that represent additional information about each LO. All this information can be very useful in order to select the best and most interesting LOs for the group.

The DELPHOS system provides the following additional information for each recommended LO in the first 8 columns: the format of the LO; a description of the LO and a button to add/select it to a group; a score calculated automatically through a hybrid filtering system based on weights (Zapata et al., 2013); a short explanation about the reasons why this particular LO has been recommended; other related and similar LOs; and the users who have downloaded and evaluated this LO.

DELPHOS also provides several links (see the last two columns in Fig. 7) to a review form for evaluating the recommended LOs, and to the next collaborative options:

- *To rate*. All the group members can rate the LOs according to the general interests and needs of the group (see Fig. 8). This rating is represented by a Likert scale of five stars (: Not recommended, : Poorly recommended, : Recommended, : Well recommended, : Extremely well recommended).
- *To tag and comment*. Group members can add one or more tags to LOs, as well as adding personal comments or additional information to them.

Finally, DELPHOS has another panel (see Fig. 9) from which group members can perform more group activities:

- *Using communication tools* (see Fig. 9 at top). Group members can hold a synchronous online discussion (chat) or an asynchronous discussion (e-mail) in order to arrive at a group consensus about each LO.
- *Setting the social level between group members* (see Fig. 9 at top). Group members must set/rate the level of the social relationship that they have with the other members of the group. This rating is represented by a Likert scale of five icons (: null; : low; : normal; : high; : very high).
- *Visualising and rating LOs* (see Fig. 9 at centre). Group members can visualise all LOs that have been added to the group by other members, and they can also rate them if they have not done so previously. This rating is represented by the same Likert scale and can be visualised by all members of the groups.
- *Arriving at a final consensus* (see Fig. 9 at centre and right). The goal of this final rating is to involve all the group members in coming to a consensus about each LO according to the general group interests. This rating is represented by a Likert scale of five stars (: not interesting; : low interest; : interesting; : very interesting; and : extremely interesting). All the group members can see this value but only the group mediator can set or change it (see “Final decision” column in Fig. 9).

DELPHOS also provides the group mediator with a different way to automatically obtain the final rating for each LO without

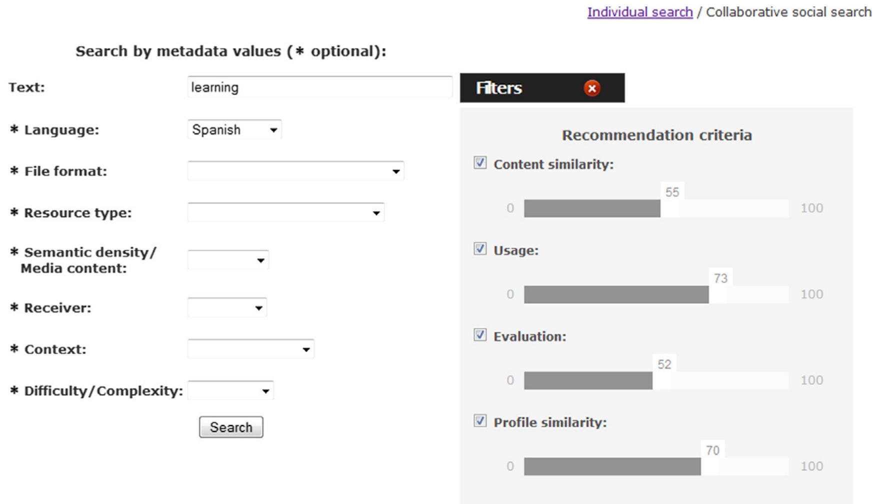

Fig. 6. Example of LOs search.

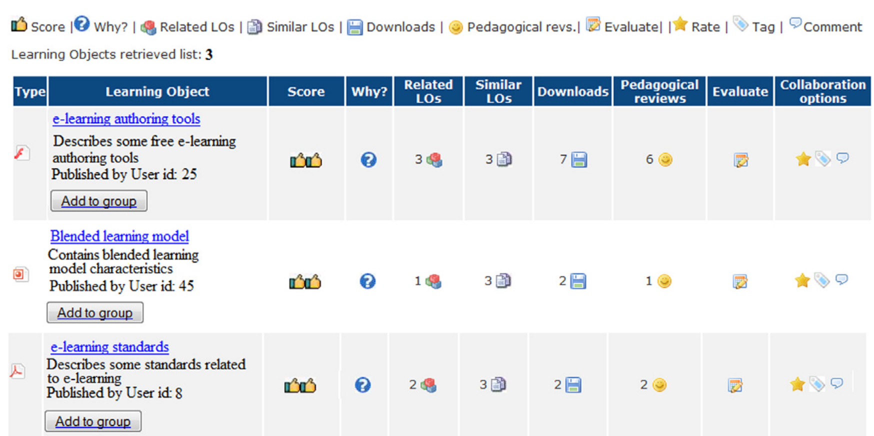

Fig. 7. Example of a list of recommended LOs in the collaborative search mode.

Table 2
Icons' description associated with the interface of recommended LOs in a collaborative search.

Icons	Name	Description
	Score	Calculation based on a 5-thumbs scale.
	Why?	Shows a short explanation about why this particular object has been recommended.
	Related LOs	Shows a list of the most downloaded objects by users who have also downloaded this LO.
	Similar LOs	Shows a list of the most similar objects according to IEEE-LOM metadata.
	Downloads	Shows how many users have downloaded this LO.
	Pedagogical reviews	Shows how many users have evaluated this LO.
	To evaluate	User can perform a pedagogical evaluation of this LO.
	To rate	User can assign a rating on a 5-star scale.
	To tag	User can add more information to the LO metadata.
	To comment	User can add comments about this LO.

needing to hold a traditional in-person meeting or online meeting with the group members. The panel of group activities allows only the group mediator to use voting aggregation strategies, and they can manage these in two different ways (see Fig. 9 at down). First, mediators can use a one-by-one strategy (see "To select aggregation strategy" list box in Fig. 9) in order to see each LO's effect on the ratings, depending on the selected aggregation strategy (see "Aggregation strategy" column in Fig. 9). The current version of DELPHOS provides the next aggregation strategies: the eight voting aggregation methods previously described in Table 1, plus three variations of the average method in which we propose to give more weight to users with more activity, more social relationships and more experience. Second, the group mediator can use the button "To recommend the best", which directly suggests the most appropriate aggregation method for a group based on its descriptors (size, activity, experience, socialness and dissimilarity). This selection is done by using the meta-learning process described in Section 3 and shown in Fig. 3. Finally, the mediator of the group can use the rating shown in the "Aggregation strategy" column to set the final rating in the "Final decision" column instead of having to hold a meeting to arrive at a consensus.

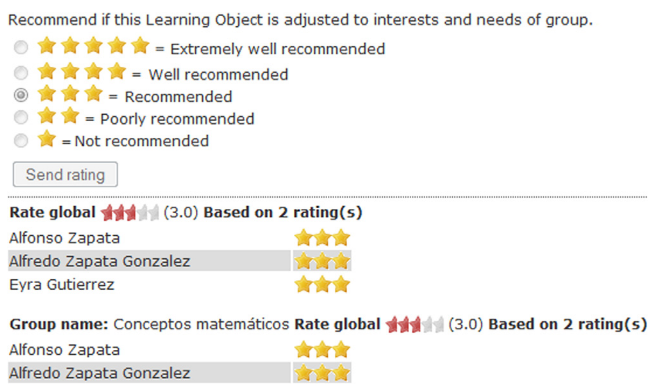

Fig. 8. Rate learning objects panel.

5. Experimental work

We carried out an experiment to validate our proposal of group recommendation approaches for searching for LOs. We sent invitations, without using any incentive, to all instructors and final-year students of the School of Education of the Autonomous University of Yucatan in Mexico to participate in the experiment. Only 75 users accepted our invitation: 27 professors or university teachers at different levels (assistant, associate and full) and 48 final-year students. Table 3 shows the distribution of all these

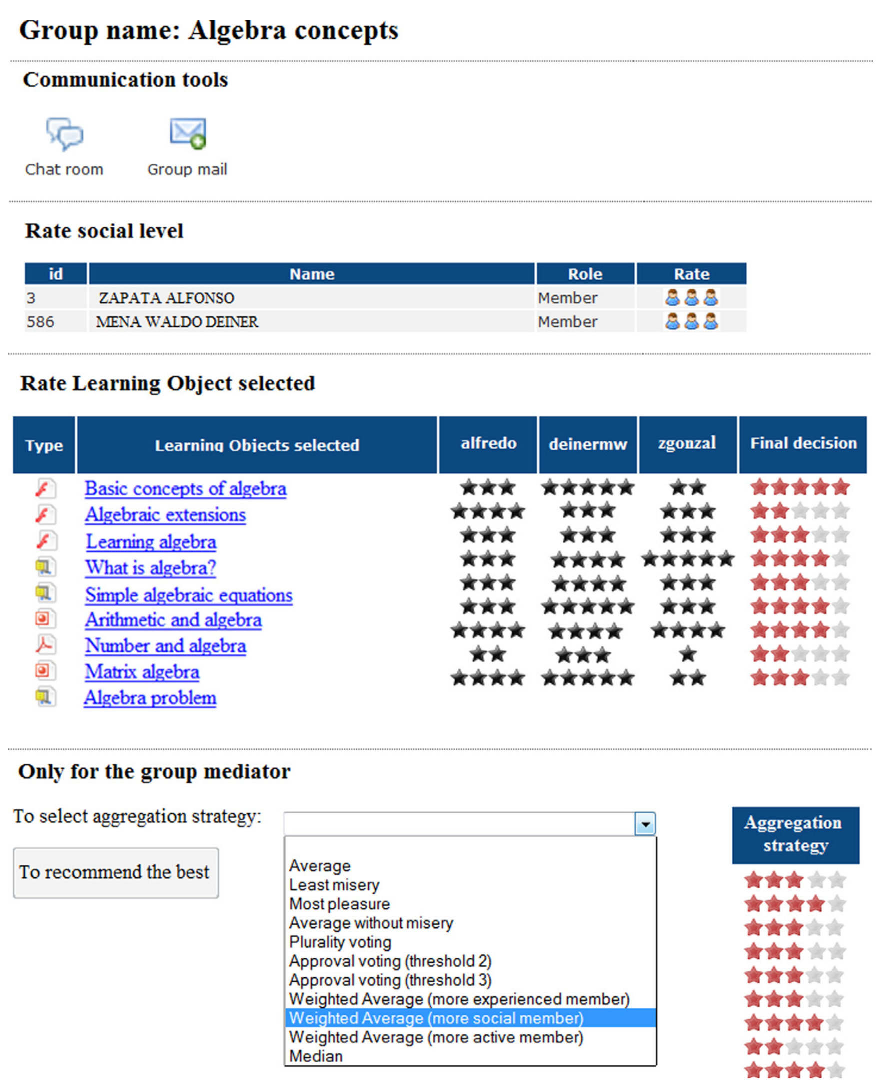

Fig. 9. The panel of group activities.

Table 3
Research area and higher-level programmes user distribution.

Research area	Numbers of professors
Curriculum and Instruction	10
Administration and Educational policies	8
Psychology and Education	9
Higher Education Programme	Numbers of student
Bachelor's Degree in Education (in-person modality)	14
Master in Educational Innovation (in-person modality)	10
Master in Educational Innovation (online modality)	24

Table 4
Group distribution by size and characteristics levels (Mean ± Desvt).

Number of groups	Group size	Activity	Experience	Social	Dissimilarity
20	2	2.06 ± 0.72	2.68 ± 0.50	2.51 ± 1.09	2.90 ± 0.21
15	3	2.09 ± 0.55	2.90 ± 0.57	2.42 ± 1.10	2.89 ± 0.18
10	4	2.07 ± 0.56	2.75 ± 0.69	2.01 ± 1.02	2.92 ± 0.10
5	5	2.00 ± 0.83	2.68 ± 0.71	2.99 ± 1.05	2.96 ± 0.13

users by research areas and higher education programmes. Table 3 shows that there are users with different levels of experience in education, from students who have zero or a very low level of experience up to professors who have a higher level of experience, depending on their position and their number of years as an instructor.

Before starting the experiment, we organised an in-person fast tutorial (1 h in duration) in a computer lab in order to provide a practical introduction to DELPHOS systems for all the new users. We also provided an online quick-start guide to DELPHOS for all online students or instructors who could not attend the tutorial. We then sent an e-mail to all users to explain what the experiment consisted of, which read as follows:

1. Users will manage their own time to access the system but they will be automatically assigned to different groups that will be directed by a group moderator.
2. Each member of the group will have to search for and select LOs about the specific theme/topic/issue established by the moderator.
3. Each member of the group will have individually to rate all the LOs selected by the group.
4. Finally, each group will have to come to a consensus about the final rating assigned to all selected LOs.

We defined a total of 50 different groups of instructors and students of different typologies. To this end, we have created groups of different sizes (see Table 4). Our objective in the groups' distribution was to have a wide variety of group typologies. The 50 groups were shaped not only of different sizes but also with different levels of experience in education, different levels of social relationships and different levels of activity. Some groups were created starting with an instructor from the same area, other groups with instructors from different areas, others with students from the same programme, others with students from different programmes, and others with both instructors and students. Finally, we decided that students and instructors could belong to different groups at the same time during this experiment.

Afterwards, the role of mediator was assigned to the eldest member in each of the groups. First, the mediator of each group was in charge of establishing and announcing the concrete theme of the group search. Thus, each mediator was free to specify the search topic for each group that he or she managed, which could have been based on a common interest or on the study programme to which most of the group members belonged. For example, the mediator of group 1 established a search of LOs related to e-learning and using the default values of the weights of the filtering criteria provided by DELPHOS. Once the search parameters were defined, the mediators of the groups established two maximum time periods to complete the next search and rating tasks:

- *To search and rate*. A maximum time of one week was established in order that all the members of the group could perform the search for the relevant LOs for the corresponding group. In addition to adding LOs to the corresponding groups, the members had to set their personal ratings/scores.
- *To get a consensual final rating*. A maximum period of one more week was established in order to get the final consensual rating for each of the aggregated LOs by all the members of the group. This final rating is assigned by the mediator of the group.

After two weeks of their searching and rating, we obtained the final ratings of the 50 groups through consensus or through a democratic process among all the members of each group. Next, with all of this information, we validated our proposal in three steps: 1) evaluating the rating aggregation methods, 2) selecting an aggregation method, and 3) evaluating the system usability.

5.1. Evaluation of rating aggregation methods

In a first step, we have evaluated the performance of 11 implemented rating aggregation strategies in DELPHOS over the 50 groups. In order to do this, we have used the RMSE (Root-Mean-Square Error) of each aggregation method in each group. This is a frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed (Hyndman and Koehler, 2006). Its values vary from zero to infinity, where smaller RMSE values indicate better predictions, and they are obtained using equation

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (X_{1,i} - X_{2,i})^2}{n}} \tag{2}$$

where

- X_1 = The real value or final LO rating that has been obtained in a consensual way by all the members of a group.
- X_2 = The predicted value or estimated LO rating, using a rating aggregation method.
- n = The total number of ratings or LOs added to a group.

Table 5
Aggregation methods ranking in group 1.

Ranking	Aggregation methods	RMSE
1	Weighted average (more active user)	0.23299295
2	Median	0.26726124
3	Average	0.27548658
4	Weighted average (more social user)	0.29760952
5	Weighted average (more experienced user)	0.30705979
6	Plurality voting	0.53452248
7	Least misery	0.70710678
8	Most pleasure	0.75592895
9	Fairness	0.75592895
10	Approval voting	0.96362411
11	Average without misery	1.65156852

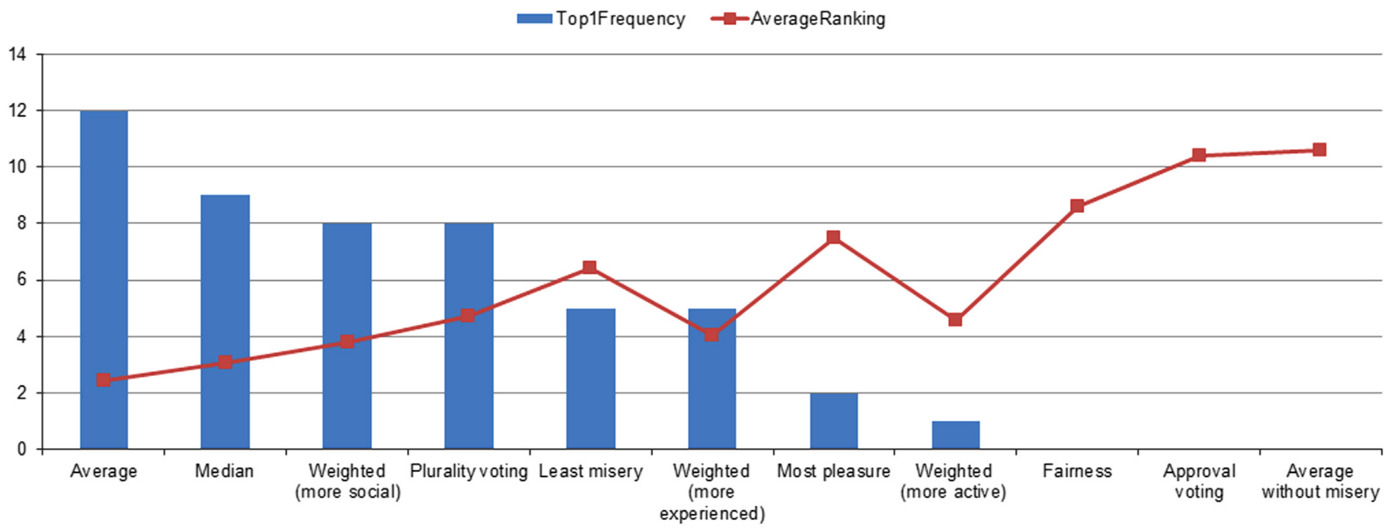

Fig. 10. Top-1 Frequency and Average Ranking of aggregation methods in the 50 groups.

RMSE provides us with the error value committed for each aggregation method with respect to the actual or final rating. Thus, the lower the RMSE value, the better the aggregation method. For example, Table 5 shows the rankings obtained for the aggregation methods used in group 1.

Starting from this ranking, we have used two evaluation measures. The first evaluation measure used is the Top-1 Frequency (Wang et al., 2008), which is the frequency or number of times that an aggregation method has obtained the first position in the rankings of the 50 groups, and is obtained by equation

$$\text{Top1Frequency} = \sum_{i=1}^{N} BestMethod_i \tag{3}$$

where

- $BestMethod_i$ = Return 1 is the method that has obtained the first position in the ranking of group i, or 0 in other cases.
- N = Number of groups.

The second measure is the Average Ranking (Konstan et al., 1997), which is the average position that an aggregation method occupies in the rankings of the 50 groups, and is obtained by equation

$$\text{Average Ranking} = \frac{\sum_{i=1}^{N} PositionMethod_i}{N} \tag{4}$$

where

- $PositionMethod_i$ = Position that a method has obtained in the ranking by group i.
- N = Number of groups.

Fig. 10 shows the values of Top-1 Frequency and Average Ranking for the 11 aggregation methods used by the 50 groups.

As shown in Fig. 10, the best aggregation method was Average, due to the fact that it had the highest frequency (Top-1 Frequency) and the lowest average position (Average Ranking). Other aggregation methods that also obtained good results with the 50 groups are median, weighted average (more social user) and Plurality voting; followed by least misery, weighted average (more experienced user), most pleasure and weighted average (more active user) in this order. Finally, the worst aggregation methods, which were determined due to the fact that they were never selected in the first position in the rank although they had the highest average positions, are Fairness, Approval voting and Average without misery.

However, these results change if we use other groups of users or even if we select only certain groups based on their descriptors/characteristics. That is to say, Top-1 Frequency and Average Ranking values of the aggregation methods vary depending on the characteristics of the groups used. Thus, some aggregation methods can work better with some types of groups and worse with others. In order to demonstrate this fact, we have carried out two tests. In the first test, we have created several datasets by selecting, starting with our 50 groups, only the groups of the same size (2, 3, 4 and 5 members) in four different datasets. We have then calculated again the performance (Top-1 Frequency and Average Ranking) of all the aggregation methods in each one of these datasets (see Fig. 11).

As we can see in Fig. 11, the best aggregation method is not always Average, and it changes with the different sizes of groups. In fact, the Average method does not obtain the best Top-1 Frequency outright in any dataset; rather, it shares the first position with other methods or it obtains the second position in the ranking. Also, the Average method only obtains the best Average Ranking in two datasets, but is equal or lower than other methods in the other two datasets. Therefore, there are some other aggregation methods that have shown a better performance than Average for certain group sizes. For example, in small groups with only two or three members, methods (such as Plurality voting and Median) that facilitate observing the tendency in rating results obtain the best performance. In contrast, in big groups with four or five members, methods (such as Weighted average for more social or more experienced users) which establish a greater weight to ratings of a particular member obtain the best performance.

In the second test, we have created other different datasets by grouping the 50 groups using a clustering algorithm. The objective is to group the most similar groups together in each dataset, depending on their characteristics/descriptors. In order to do this, we have used the EM (Expectation Maximisation) algorithm (Dempster et al., 1977), because it is not necessary to indicate the number of subgroups or clusters to be found. EM discovered two different clusters over the 50 groups, and the mean and standard deviation values of their centroids (representing the most typical group in each cluster) are shown in Table 6.

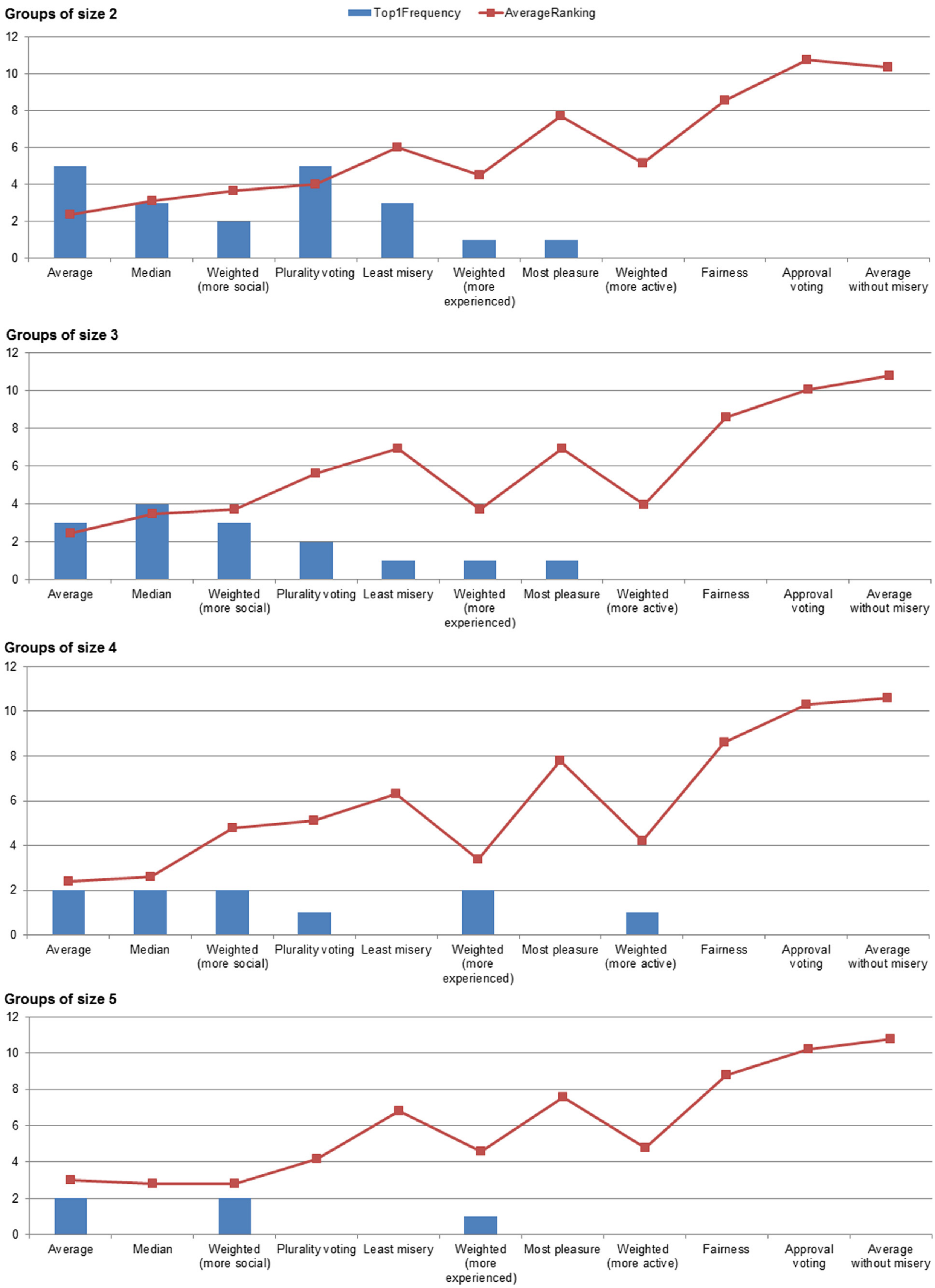

Fig. 11. Top-1 Frequency and Average Ranking of aggregation methods by group size.

As is shown in Table 6, cluster 1 contains smaller groups with less activity, experience, social relationship and dissimilarity than the larger groups of cluster 2.

Next, we have again calculated the performance (Top-1 Frequency and Average Ranking) of all the aggregation methods in each of these two datasets (see Fig. 12).

As we can see in Fig. 12, in cluster groups 1, the Plurality voting method obtained the best Top-1 Frequency and the Median method the best Average Ranking. In cluster groups 2, the Median obtained both the best Top-1 Frequency and Average Ranking. In both clusters, the Average method obtained the second best Top-1 Frequency and Average Ranking.

In summary, the Average aggregation method obtained the best performance with the 50 groups (see Fig. 10); however, when we have used datasets that contained only groups of similar sizes (see Fig. 11) or with similar characteristics (see Fig. 12), the best performance was obtained by different aggregation methods.

5.2. Selection of an aggregation method

In a second step, we want to predict the most appropriate aggregation method to use with a new group, based on the characteristics of the group members and the previous rating of similar groups.

Table 6
Centroid values of each cluster.

Attribute	Cluster groups 1	Cluster groups 2
Size	2.3019 ± 0.5787	3.1657 ± 0.92
Activity	1.8006 ± 0.6127	2.5427 ± 0.3549
Experience	2.436 ± 0.3956	3.3938 ± 0.1813
Social	2.2196 ± 0.8665	2.8527 ± 1.2761
Dissimilarity	1.9456 ± 0.1739	2.8438 ± 0.1438

In order to carry out this step, we have created a metadata set that contains both the characteristics/descriptors of all the groups as well as the best aggregation methods for each group (the first aggregation method in the ranking of RMSE). This metadata set includes the descriptors or characteristics of the 50 groups (see Table 7). For each characteristic, we have used both the mean (average between the members) and the variance (deviation values with respect to the mean).

Starting from this metadata set, it is possible to predict which is the best aggregation method that a new group must use. This is a classification in which the class or attribute to predict is precisely the aggregation method that obtains the best ranking. To this end, we have used different classification algorithms provided by the WEKA software (Witten and Frank, 2005), which is one of the most popular and most used tools for data mining. We have selected a representative number of the best known classification algorithms available in WEKA: JRip (implementation of RIPPER algorithm), J48 (implementation of C4.5 algorithm), NaiveBayes-Simple (implementation of Bayes classifier), SMO (implementation of support vector classifier) and IBk (implementation of KNN or Nearest Neighbours algorithm). We have executed the previous five classification algorithms using their default parameter values and 10-fold cross-validation. This means that the dataset is divided randomly into 10 disjoint subsets of equal size in a stratified manner (maintaining the original class distribution). The algorithm is executed 10 times, and in each repetition, one of the 10 subsets is used as the test set and the other nine subsets are combined to form the training set. The classification performance

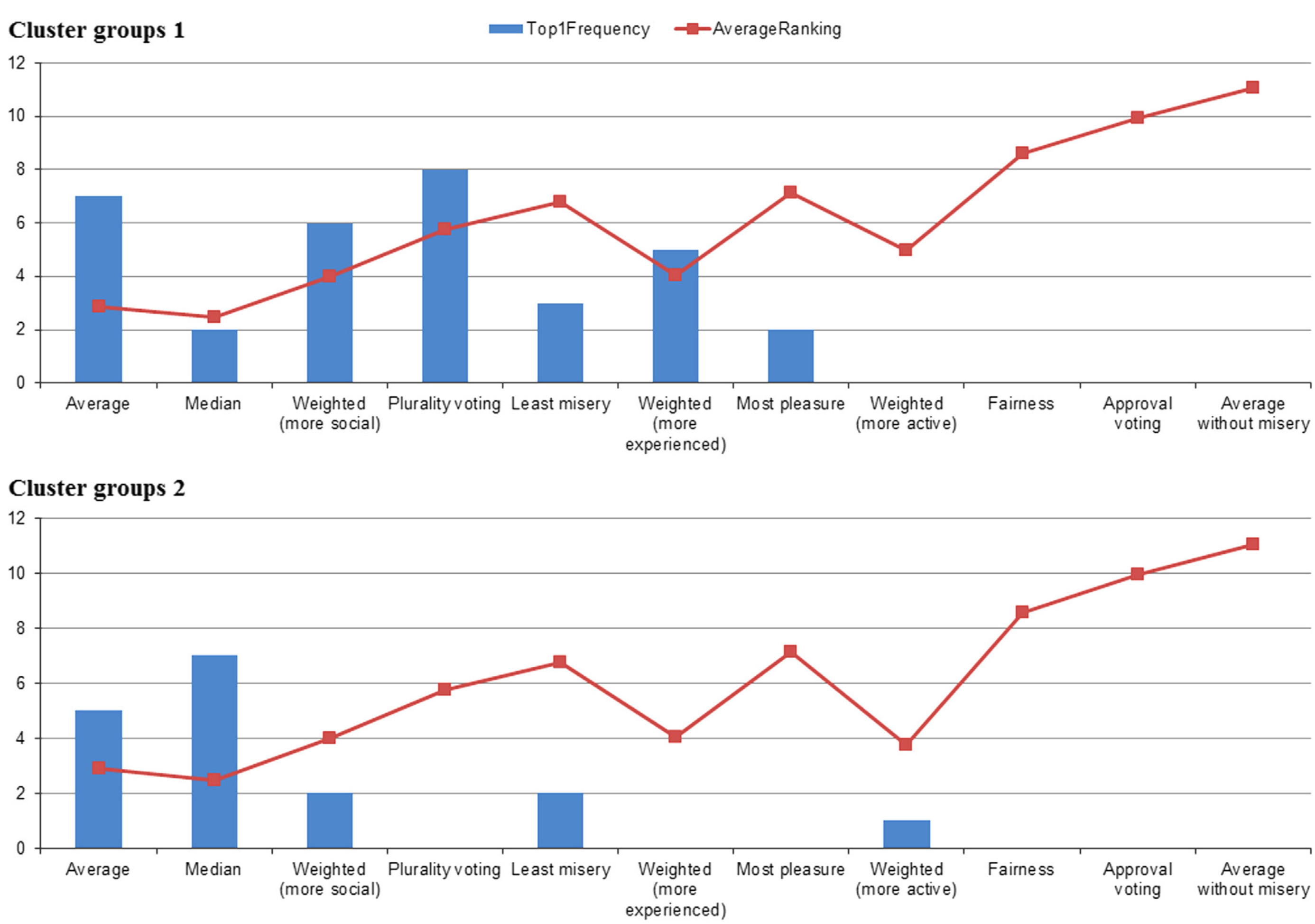

Fig. 12. Top-1 Frequency and Average Ranking of the aggregation methods in cluster groups 1 and 2.

Table 7
Groups' size and descriptors.

ID	Size	Activity	Experience	Social	Dissimilarity
1	5	3 ± 1.5	3.4 ± 0.77	4.3 ± 0.43	2.98 ± 1.61
2	4	2.25 ± 1.58	3.5 ± 0.4	4.5 ± 0.27	2.80 ± 1.17
3	3	2.33 ± 0.33	3.5 ± 0.27	5 ± 0	3 ± 0.84
4	3	2.33 ± 2.33	3.5 ± 0.45	4.16 ± 0.16	2.83 ± 1.17
5	2	2.5 ± 0.5	3.25 ± 0.78	5 ± 0	2.95 ± 1.43
6	5	3 ± 2.5	3.5 ± 0.68	3.18 ± 2.29	2.84 ± 1.40
7	4	2.5 ± 1.66	3.56 ± 0.26	2.58 ± 3.53	2.80 ± 1.39
8	3	2 ± 1	3.41 ± 0.44	1.66 ± 0.66	2.96 ± 1.46
9	3	2.66 ± 4.33	3.25 ± 0.93	1.5 ± 0.7	2.66 ± 1.05
10	3	2.33 ± 2.33	3.25 ± 0.57	1.83 ± 0.16	2.54 ± 1.88
11	2	3 ± 2	3.62 ± 0.55	3.5 ± 0.5	2.54 ± 1.04
12	2	3 ± 2	3.37 ± 0.55	2.5 ± 0.5	2.91 ± 1.38
13	3	2.33 ± 2.33	3 ± 1.09	1.83 ± 0.96	2.86 ± 0.80
14	2	1.66 ± 0.33	2.83 ± 1.24	1.33 ± 0.26	2.58 ± 1.56
15	4	2 ± 0.66	3.12 ± 1.18	1.66 ± 1.15	3.08 ± 1.39
16	3	2.66 ± 1.33	3.25 ± 0.93	1.66 ± 0.66	2.90 ± 0.86
17	3	1.66 ± 0.33	2.33 ± 0.96	1.33 ± 0.66	3.19 ± 1.13
18	3	1.33 ± 0.33	2.33 ± 1.15	1.75 ± 2.25	2.88 ± 1.58
19	2	2.5 ± 0.5	2.75 ± 1.35	1.5 ± 0.5	3.04 ± 1.34
20	4	2.5 ± 4.5	2.75 ± 1.35	1.5 ± 0.5	2.87 ± 1.15
21	2	3 ± 2	2.62 ± 1.12	1.5 ± 0.5	2.73 ± 1.40
22	3	2 ± 1	3.75 ± 0.38	2.16 ± 2.16	2.78 ± 1.48
23	2	2.66 ± 0.33	3.13 ± 0.98	2.66 ± 1.86	2.80 ± 1.47
24	2	2.5 ± 0.5	2.87 ± 1.26	2.5 ± 0.5	3.25 ± 1.15
25	4	3 ± 2	3.5 ± 1.14	1 ± 0	2.87 ± 0.72
26	2	3 ± 2	2.87 ± 0.98	1 ± 0	3 ± 1.04
27	3	3.33 ± 2.33	2.66 ± 0.96	1.83 ± 1.36	2.79 ± 1.43
28	2	2.5 ± 0.5	3 ± 1.14	1 ± 0	3 ± 1.65
29	4	1.5 ± 0.5	1.87 ± 0.41	1 ± 0	2.96 ± 1.47
30	2	1.5 ± 0.5	1.75 ± 0.21	2.5 ± 0.5	2.88 ± 1.62
31	3	1.5 ± 0.5	1.75 ± 0.21	4 ± 0	2.79 ± 1.47
32	4	1.5 ± 0.5	1.87 ± 0.41	1.5 ± 0.5	2.89 ± 1.80
33	2	1 ± 0	2 ± 0.57	2.5 ± 0.5	2.57 ± 1.45
34	2	1.5 ± 0.5	2 ± 0.57	1.5 ± 0.5	2.80 ± 1.44
35	4	1.5 ± 0.5	1.87 ± 0.41	2 ± 0	2.85 ± 1.68
36	2	1.5 ± 0.5	2.25 ± 0.5	3.5 ± 0.5	2.87 ± 1.24
37	3	1.5 ± 0.5	2.62 ± 1.12	2.5 ± 0.5	3 ± 1.52
38	2	1 ± 0	2.62 ± 0.55	2.5 ± 0.5	2.84 ± 1.49
39	5	1 ± 0	2.25 ± 0.5	3.5 ± 0.5	3.10 ± 1.43
40	5	1.5 ± 0.5	2.25 ± 1.07	1.5 ± 0.5	2.96 ± 1.07
41	2	1.5 ± 0.33	2.43 ± 0.92	1.83 ± 0.33	3.08 ± 1.46
42	4	1.25 ± 0.25	2.87 ± 1.58	1.91 ± 0.26	3.05 ± 1.14
43	3	2 ± 1	2.33 ± 0.96	2.66 ± 1.06	3.15 ± 1.34
44	5	1.5 ± 0.5	2 ± 0.57	2.5 ± 0.5	2.75 ± 1.30
45	2	2 ± 0	2.75 ± 1.35	2.5 ± 0.5	2.78 ± 1.06
46	2	1 ± 0	2.37 ± 0.83	4.5 ± 0.5	3.35 ± 1.35
47	2	2.5 ± 0.5	2.12 ± 0.98	3 ± 2	3.04 ± 1.34
48	2	1.5 ± 0.5	3.12 ± 1.83	3.5 ± 0.5	3 ± 1.40
49	4	2 ± 0	2.62 ± 0.55	2.5 ± 0.5	3.03 ± 1.14
50	3	1.5 ± 0.5	2.62 ± 0.83	2.5 ± 0.5	3.14 ± 1.16

values are calculated by averaging the results obtained in the training sets.

In order to evaluate the classification performance and to determine which is the best algorithm for each group, we have used two measures that have previously been used to evaluate classification algorithm recommendation methods (Song et al., 2012). The first is called ARE (Average Recommendation Error) and it measures the average error of the current recommendation (predicted aggregation method) regarding the best and the worst recommendation (best and worst aggregation methods from the list of methods ordered from the lowest to the highest RMSE), as expressed in equation

$$\text{ARE} = \frac{1}{|N|}\sum_{i=1}^{|N|}\frac{ERRORnow_i - ERRORworst_i}{ERRORbest_i - ERRORworst_i} \tag{5}$$

where

- N=Number of groups.
- $ERRORnow_i$=The RMSE value of the aggregation method, predicted by the algorithm for the group.
- $ERRORworst_i$=The RMSE value of the worst aggregation method for the group (method located in the last place in the ranking).
- $ERRORbest_i$=The RMSE value of the best aggregation method for the group (method located in the first place in the ranking).

The second measure is the Reciprocal Average Hit Rate, also known as Mean Reciprocal Rank (MRR) (Bian et al., 2008), which measures the median position occupied by the method currently predicted for each of the groups in the complete list of methods ordered by RMSE, and calculated through equation

$$\text{MRR} = \frac{1}{|N|}\sum_{i=1}^{|N|}\frac{1}{rank_i} \tag{6}$$

where

- N=Number of groups.
- $rank_i$=The rank or position of a method in the ordered list of group i.

Fig. 13 shows the values of the ARE and the MRR of the five classification algorithms. As we can see in Fig. 13, IBk was the best classification/prediction algorithm (followed by NaiveBayes and J48) because it obtained the lowest value of Average Recommendation Error and the lowest value of Mean Reciprocal Rank. So, since the algorithm IBk achieved the best results, we integrated it into the DELPHOS interface to automatically recommend it (as the best aggregation method of the most similar group or nearest neighbours) to every new group as the best method for rating all the LOs added to the group. In this way, the moderator of the group would use the recommended aggregation method obtained (i.e., the IBk algorithm) instead of having to conduct the traditional consensual decision process.

5.3. Evaluation of the system usability

We have evaluated the system usability in two phases. The first phase was just after all participants (75 users who participated in the previous experiments) completed an online questionnaire in order to give their personal opinion about the general usability of our system (see Table 8). We used the Computer System Usability Questionnaire (CSUQ) (Lewis, 1995), which was developed by IBM and is composed of a scale of 19 items or questions. It is rated through a seven-point scale from 1 (strongly disagree) to 7 (strongly agree) and Not Applicable (NA). The general degree of usability of the system was obtained by averaging the answers of all users on one single value between 0% and 100%.

The results of the CSUQ questionnaire (see Table 8) show that the users have a good opinion about the functionalities provided by the DELPHOS system, because they obtain a value of 77.48% as the average score. In general, they consider that the system is easy to use and it facilitates greatly all the actions provided by our system.

The second phase started some weeks later than the first one, but we also needed two months more in order to finish it. The reason for carrying out this second phase was that after analysing the results of the CSUQ questionnaire, we became aware that our evaluation, while positive, was lightweight. That is, CSUQ was well known as a general usability questionnaire, but it did not show us any specific information regarding the user's perception of collaborative searches and the recommended aggregation technique. Thus, we decided to contact all the participants in the experiment again in order to request the completion of two more specific

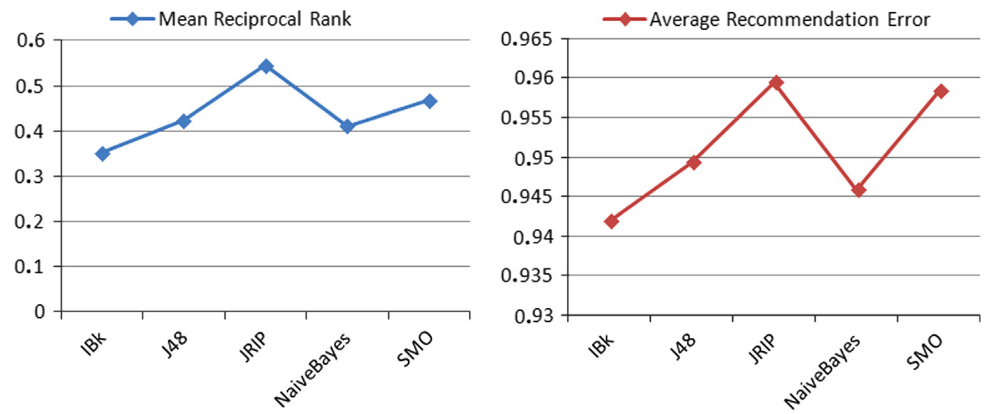

Fig. 13. Average Recommendation Error and Mean Reciprocal Rank obtained by the 5 classification algorithms.

Table 8
Results of CSUQ questionnaire.

Question		Average
1.	Overall, I am satisfied with how easy it is to use the system.	6.10
2.	It was simple to use the system.	5.96
3.	I can effectively complete my work using the system.	5.40
4.	I am able to complete my work quickly using the system.	5.80
5.	I am able to efficiently complete my work using the system.	5.84
6.	I feel comfortable using the system.	5.36
7.	It was easy to learn to use the system.	5.67
8.	I believe that I became productive quickly using the system.	4.94
9.	The system gives error messages that clearly tell me how to fix problems.	4.08
10.	Whenever I make a mistake using the system, I recover easily and quickly.	4.85
11.	The information (such as online help, on-screen messages, and other documentation) provided with the system is clear.	5.03
12.	It was easy to find the information I needed.	5.48
13.	The information provided for the system is easy to understand.	5.26
14	The information is effective in helping me complete the tasks and scenarios.	4.70
15	The organisation of information on system screens is clear.	5.24
16.	The interface of the system is pleasant.	5.66
17.	I like using the interface of the system.	6.16
18.	The system has all the functions and capabilities I expect it to have.	5.90
19.	Overall, I am satisfied with the system.	5.63
Usability		**77.48**

Table 9
Description of the Morris questionnaire about satisfaction in collaborative search.

Attribute	# Question	Description
Informational	Q1	About the collaborative search (CL) of learning objects: the quality of the found learning object
Collaborative	Q2	About the CL: easy to work with other users using DELPHOS

Table 10
Results of the Morris questionnaire.

# Question	Very Satisfied (%)	Satisfied (%)	Neutral (%)	Dissatisfied (%)	Very Dissatisfied (%)
Q1	19.05	66.67	14.29	0.00	0.00
Q2	33.33	42.86	23.81	0.00	0.00

Table 11
Description of the perceived usefulness section of the ResQue questionnaire.

Attribute	# Question	Description
Utility	Q1	About the Recommender System (RS) to select the best aggregation strategy for a specific group: This RS helped me find the ideal aggregation strategy
Utility	Q2	About the RS: This system influenced my final decision
Easiness	Q3	About the RS: Using this system to find an ideal aggregation strategy is easy
Easiness	Q4	About the RS: Finding an final decision with the help of this system is easy

questionnaires. The first questionnaire (Morris, 2013) was focused on evaluating the satisfaction in the collaborative search experience and was sent to all 75 users. The second one (Pu et al., 2011) was focused on evaluating the usefulness of the recommendation of aggregation techniques and was sent only to the 50 mediators of the groups, because the mediators were the only users who viewed and used the voting aggregation strategies in DELPHOS. The problem was that most of the participants required much time to send us the filled questionnaire, and even after several reminders, some of them never answered us. In the end, we only obtained responses from 42 users and 21 mediators.

The first questionnaire was Morris's questionnaire (2013) about users' satisfaction when doing collaborative Web searches. This questionnaire (see Table 9) had only two questions, with a five-point Likert scale (from Very Satisfied to Very Dissatisfied) that we have adapted in order to use the same scale as that of the ResQue questionnaire (Pu et al., 2011). We have also slightly adapted the wording or text of the two questions in order to be much more specific about the collaborative search of LOs using DELPHOS. Question 1 was about the informational quality of the learning object found by the group, and Question 2 was about the collaborative aspects of their search (i.e., how easy it was to work with other users using DELPHOS).

Table 10 summarises the results obtained from the Morris questionnaire. Accordingly, 85.72% users reported satisfaction with the informational outcome and only 14.29% users reported a neutral opinion in Question 1. For question two, 76.19% users reported satisfaction with the ease of collaboration and 23.81% offered a neutral opinion. Although these results demonstrate a positive overall assessment, they also show us that we can improve some collaborative aspects in DELPHOS in order to make it easier to work with other users.

We also requested that users give their personal comments and suggestions in free-form mode with regard to their collaborative search experience. A common suggestion from the users was about providing more facilities related to group awareness of mutual activities. For example, one user said, “I would like to be notified by mail when a team member selects an object [such as x].” Another noted, “I think that it is very useful to have a progress bar that shows the completion percentage of the collaborative searching.” In this line of proposing novel system features, another user said “it would be useful to incorporate an indicator or mark [in] the learning objects [that] have been selected by any of the members of the groups to which I belong”. Finally, a recurring comment discussed the possibility of having a mobile version of the system that exploits the characteristics of different types of mobile devices.

The second questionnaire was oriented to validate the users' perceived usefulness of the voting aggregation strategies and the recommendations provided to the group mediators. We used Section 5 of the Recommender Systems' Quality of User Experience (ResQue) questionnaire (Pu et al., 2011). This section of the ResQue questionnaire has four questions (see Table 11) that users must rate using a five-point Likert scale from Very Satisfied to Very Dissatisfied. We have slightly adapted the wording or text of the four questions in order to be much more specific about the aggregation strategy of our recommender system.

Table 12
The results of the ResQue questionnaire.

# Question	Very Satisfied (%)	Satisfied (%)	Neutral (%)	Dissatisfied (%)	Very Dissatisfied (%)
Q1	38.10	52.38	9.52	0.00	0.00
Q2	23.81	38.10	38.10	0.00	0.00
Q3	38.10	52.38	9.52	0.00	0.00
Q4	57.14	42.86	0.00	0.00	0.00

Table 12 summarises the results obtained from the ResQue questionnaire. An average of 76.19% (Q1 and Q2) mediators reported satisfaction with the utility of finding the ideal aggregation strategy and of making a final decision using the said strategy, and an average of 95.24% (Q3 and Q4) mediators reported satisfaction with the easiness of using this RS. These are positive overall results, which show the highest value in Q4 and the lowest in Q2. We think that differences between these two values regarding utility and easiness could be due to two facts: one, that the mediators found it easy to use our recommender system; two, that it nevertheless did not provide any additional explanation about how the best aggregation strategy was selected or how the final ratings were generated. Therefore, the mediators continued to prefer to use the traditional consensus regarding the final rating. In fact, some mediators mentioned this in their free-form suggestions and comments. For example, one mediator commented, “The recommendation button is very useful, and I think that the manual selection of aggregation strategy is unnecessary.” However, another mediator said, “It is a good thing that one can use the recommendation button because a final rate is easily obtained. However, I prefer to conduct a traditional discussion between persons instead of directly using the decision of an algorithm.”

6. Conclusions and future work

In this paper, we have proposed a model for searching, selecting and rating LOs in groups of users. We have implemented and tested this model within the environment of the DELPHOS hybrid recommendation system. Regarding the results obtained with our 50 different groups, we evaluated the performance of 11 rating aggregation methods by using Top-1 Frequency and Average Ranking measures. Our results showed that the Average method was the best when we used the data of all our groups. However, alternative methods, such as Plurality voting or the Median and Weighted average methods, obtained the best performance when using only data from groups of different sizes (2, 3, 4 or 5) or different characteristics (2 different clusters). These results imply that there is no single best aggregation method for all the cases, as has been shown in other related works (Gartrell et al., 2010; Amer-yahia et al., 2009). Therefore, group recommendation is a challenging problem due to the dynamics and diversity of groups. For this reason, we have preferred not to explore or highlight our findings regarding the most appropriate aggregation strategy in each case.

However, we think that the most important contribution of our work is the proposal of a meta-learning approach for recommending the best aggregation strategy. In fact, our system can recommend in real-time to a group mediator the most appropriate aggregation method for a group, based on previous groups with similar characteristics. We have used five well-known classification algorithms for recommending the best aggregation method. ARE and MRR measures were used to evaluate the performance of these classification algorithms. Results showed that the IBk algorithm performed better than the other algorithms, so we have implemented it into the DELPHOS interface as a button in order to automatically recommend the best rating aggregation method for the mediator, according to a group's descriptors. In this way, mediators will not need to organise a discussion among group members to reach a consensus for each LO. Finally, we evaluated the usability of our implementation using several questionnaires. CSUQ showed that all 75 users considered that, in general, the

DELPHOS system was easy to use and facilitated the actions of searching for, retrieving and rating LOs. Morris's questionnaire showed a high satisfaction of the collaborative search experience by 42 users. Lastly, the ResQue questionnaire showed a high perceived usefulness of the voting aggregation strategies and the recommendations provided by 21 group mediators.

Finally, this work has some limitations in that we only used the AGORA repository as an experimental issue. That is, we limited the search of LOs to this repository for this experiment. However, DELPHOS allows the search to be extended through a federated search with other repositories. Another limitation is the compressibility and acceptability of our group recommender. Currently, DELPHOS does not provide any explanation about how the final rating is generated using the best aggregation strategy. This additional information could provide credibility and improve users' understanding of received recommendations (Jameson and Smyth, 2007). If this were in place, mediators could be more motivated to use the recommended aggregation strategy without needing to conduct a discussion among group members. A different limitation in our usability evaluation is the fact that we had to continue collecting data at a much later date than the original study, and it explains that not all users responded to the last two questionnaires. Thus, it is also necessary to consider certain time-related issues, such as the fact that users require a greater mental effort and a complex cognitive process for remembering the interface details once time has elapsed (Hassenzahl and Sandweg, 2004; Nisbett and Wilson, 1977). Also, web experiments frequently incur in users behaviours that affect the evaluations in a negative way, for example, the drop-out of the experiments at any time because the participation is completely voluntary, contrary to experiments controlled in a laboratory (Reips, 2000). Another aspect that we have not covered in the usability evaluation results was the required hard work of users. First, each group member had to rate all of the groups' selected learning objects. This might become burdensome or effortful once the number of LOs becomes very high. In order to resolve this problem, we want to work with incomplete data, which will not require users to rate all LOs. Second, we requested users to set their level of social contact with the other members of the group. This is a tedious task that could also produce inconsistencies among users. To solve it, we want to automatically obtain a unique value for each pair of users by means of information obtained from external data sources. Related works (Xiang et al., 2010) have shown that relationships' strength can be accurately inferred from interaction activity in online social networks such as Facebook and Twitter, or in professional/academic networks such as LinkedIn and ResearchGate. As additional future work, we want to carry out more experiments using a greater number of classification algorithms for the meta-learning stage, and a wider diversity of groups of users from different areas and domains such as engineering or health sciences, in order to obtain more generalisable results.

Acknowledgements

This research has been partially supported by the TIN2010-20395 FIDELIO project and the POII10-0133-3516 PLINIO project, JCCM, Spain; the Spanish Ministry of Science TIN-2011-22408, and the National Council of Science and Technology (CONACYT), México.

References

Aha, D., Kibler, D., Albert, M., 1991. Instance-based learning algorithms. Mach. Learn. 6 (1), 37–66.
Amershi, S., Morris, M.R., 2008. CoSearch: a system for co-located collaborative Web search. In: Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems, CHI '08. ACM, New York, NY, USA, pp. 1647–1656.
Amer-yahia, S., Roy, S.B., Chawla, A., Das, G., Yu, C., 2009. Group recommendation: semantics and efficiency. In: Proceedings of the VLDB Endowment, 2(1), pp. 754–765.
Berkovsky, S., Freyne, J., 2010. Group-based recipe recommendations: analysis of data aggregation strategies. In: Proceedings of RecSys. Barcelona, pp. 111–118.
Bian, J., Liu, Y., Gichtein, E., Hongyuan, Z., 2008. Finding the right facts in the crowd: factoid question answering over social media. In: Proceedings of the 17th International Conference on World Wide Web, pp. 467–476.
Boratto, L., Carta, S., Chessa, A., Agelli, M., Clemente, M.L., 2009. Group recommendation with automatic identification of users communities. In: Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology. IEEE Computer Society, Washington, DC, USA, pp. 547–550.
Boratto, L., Carta, S., 2010. State-of-the-art in group recommendation and new approaches for automatic identification of groups. In: Soro, A., Eloisa, V., Giuliano, A., Gavino, P. (Eds.), Information Retrieval and Mining in Distributed Environments. Springer, Berlin, pp. 1–20.
Brazdil, P., Giraud-Carrier, C., Soares, C., Vilalta, R., 2009. Metalearning: Applications to Data Mining, first ed. Springer, Berlin.
Campos, L.M., Fernández-Luna, J.M., Huete, J.F., Rueda-Morales, M.A., 2008. Managing uncertainty in group recommending processes. User Model. User Adap. 19 (3), 207–242.
Chen, Y.L., Cheng, L.C., Chuang, C.N., 2008. A group recommendation system with consideration of interactions among group members. Expert Syst. Appl. 34 (3), 2082–2090.
Cohen, W., 1995. Fast effective rule induction. In: Proceedings of the 12th International Conference on Machine Learning. Morgan Kaufmann, California, USA, pp. 115–123.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Stat. Methodol. 39 (1), 1–38.
Duch, W., 2011. Meta-learning. In: Dubitzky, W., Wolkenhauer, O., Cho, K.-H., Yokota, H. (Eds.), Encyclopedia of Systems Biology, volume 1. Springer, London.
Duda, R.O., Hart, P.E., Stork, D.G., 2000. Patter Classification, second ed. Wiley-Interscience, New York.
Gartrell, M., Xing, X., Lv, Q., Beach, A., Han, R., Mishra, S., Seada, K., 2010. Enhancing group recommendation by incorporating social relationship interactions. In: Proceedings of the 16th ACM International Conference on Supporting Group Work, GROUP '10. ACM Press, New York, NY, USA, pp. 97–106.
Golovchinsky, G., Adcock, J., Pickens, J., Qvarfordt, P., Back, M., 2008. Cerchiamo: a collaborative exploratory search tool. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '08. ACM Press, New York, NY, USA, pp. 315–322.
Gupta, A., Paul, S., Jones, Q., Borcea, C., 2004. Automatic identification of informal social groups and places for geo-social recommendations. Int. J. Mob. Netw. 2 (3/4), 159–171.
Hämäläinen, W., Vinni, M., 2011. Classiers for educational data mining. In: Romero, C., Ventura, S., Pechenizkiy, M., Baker, R. (Eds.), Educational Data Mining Handbook. CRC Press, Florida, pp. 57–74.
Hassenzahl, M., Sandweg, N., 2004. From mental effort to perceived usability: transforming experiences into summary assessments. In: Proceedings of the CHI '04 Extended Abstracts on Human Factors in Computing Systems (CHI EA '04). ACM, New York, NY, pp. 1283–1286.
Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. Int. J. Forecast. 22 (4), 679–688.
Jameson, A., Smyth, B., 2007. Recommendation to groups. In: Proceeding of the Adaptive Web Conference, pp. 596–627.
Jankowski, N., Duch, W., Grąbczewski, K., 2011. Meta-Learning in Computational Intelligence. Springer, Berlin.
Jumper Networks Inc., 2011. Jumper 2.0. Retrieved June 25, 2013 from ⟨http://en.wikipedia.org/wiki/Jumper_2.0⟩.
Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., Riedl, J., 1997. GroupLens: applying collaborative filtering to Usenet news. Commun. ACM 40 (3), 77–87.
Lewis, J.R., 1995. IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. Int. J. Hum. Comput. Interact. 7 (1), 57–78.
Manouselis, N., Drachsler, H., Vuorikari, R., Hummel, H., Koper, R., 2011. Recommender systems in technology enhanced learning. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (Eds.), Recommender Systems Handbook. Springer Press, New York, pp. 387–415.
Masthoff, J., 2011. Group recommender systems: combining individual models. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (Eds.), Recommender Systems Handbook. Springer Press, New York, pp. 677–702.
Menendez, V.H., Castellanos, M.E., Zapata, A., Prieto, M.E., Romero, F.P., 2011. Characterising the re-use of learning objects in a specialised repository. In: Proceedings of the IJCA on Design and Evaluation of Digital Content for Education, pp. 20–24.
Menéndez, V., Prieto, M., Zapata, A., 2010. Sistemas de Gestión Integral de Objetos de Aprendizaje. IEEE-RITA. Revista Iberoamericana de Tecnologías del Aprendizaje 5 (2), 56–62.
Molina, M.D.M., Romero, C., Ventura, S., Luna, J.M., 2012. Meta-learning approach for automatic parameter tuning: a case study with educational datasets. In: Proceedings of the International Conference on Educational Data Mining (EDM). Crete, Greece, pp. 180–183.

Morris, M.R., 2008. A survey of collaborative Web search practices. In: Proceeding of the 26th Annual CHI Conference on Human Factors in Computing Systems, CHI '08. ACM Press, New York, NY, USA, pp. 1657–1660.

Morris, M.R., 2013. Collaborative search revisited. In: Proceedings of the 2013 Conference on Computer-Supported Cooperative Work (CSCW 2013). ACM, New York, NY, pp. 1181–1192.

Morris, M.R., Horvitz, E., 2007. SearchTogether: an interface for collaborative Web search. In: Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology, UIST '07. ACM Press, New York, NY, USA, pp. 3–12.

Nisbett, R.E., Wilson, T.D., 1977. Telling more than we can know: verbal reports on mental processes. Psychol. Rev. 84, 231–259.

Platt, J.C., 1999. Fast training of support vector machines using sequential minimal optimization. In: Schölkopf, B., Burges, C.J.C., Smola, A.J. (Eds.), Advances in Kernels Methods: Support Vector Learning. MIT Press, Cambridge, pp. 185–208.

Popescu, G., 2013. Group recommender systems as a voting problem. In: Proceedings of the 5th International Conference on Online Communities and Social Computing, OCSC 2013. Springer, Berlin, pp. 412–421.

Prieto, M.E., Menendez, V.H., Segura, A., Vidal, C., 2008. A recommender system architecture for instructional engineering. In: Proceedings of Emerging Technologies and Information Systems for Knowledge Society. Springer, Berlin, pp. 314–321.

Pu, P., Chen, L., Hu, R., 2011. A user-centric evaluation framework for recommender systems. In: Proceedings of the Fifth ACM Conference on Recommender Systems. ACM, pp. 157–164.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Francisco.

Reips, U.D., 2000. The Web experiment method: advantages, disadvantages, and solutions. In: Birnbaum, M.H. (Ed.), Psychological Experiments on the Internet. Academic Press, San Diego, CA, pp. 89–114.

Romero, C., Ventura, S., 2013. Data mining in education. WIRES Data Min. Knowl. Discov. 3 (1), 12–27.

Romero, C., Olmo, J.L., Ventura, S., 2013. A meta-learning approach for recommending a subset of white-box classication algorithms for Moodle datasets. In: Proceedings of International Conference on Educational Data Mining (EDM). Memphis, TN, pp. 268–271.

Schell, G.P., Burns, M., 2002. Merlot: a repository of e-learning objects for higher education. e-Service J. 1 (2), 53–64.

Seeks Project, 2011. Retrieved June 26, 2013 from ⟨http://www.seeks-project.info/site/⟩.

Smyth, B., Briggs, P., Coyle, M., Mahony, M.P., 2009. A case-based perspective on social Web search. In: Proceedings of the 8th International Conference on Case-Based Reasoning: Case-Based Reasoning Research and Development, ICCBR '09. Springer, Berlin, pp. 494–508.

Smyth, B., Coyle, M., Briggs, P., 2011. On communities, collaboration, and recommender systems in personalized Web search. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (Eds.), Recommender Systems Handbook. Springer, New York, pp. 579–644.

Song, Q., Wang, G., Wang, C., 2012. Automatic recommendation of classification algorithms based on data set characteristics. Pattern Recognit. 45 (7), 2672–2689.

Stefaner, M., Vecchia, E.D., Condotta, M., Wolpers, M., Specht, M., Apelt, S., 2007. MACE-enriching architectural learning objects for experience multiplication. In: Proceedings of the Second European conference on Technology Enhanced Learning: Creating New Learning Experiences on a Global Scale, EC-TEL '07. Springer, Berlin, pp. 322–336.

Ternier, S., Verbert, K., Parra, G., Vandeputte, B., Klerkx, K., Duval, E., Ordonez, V., Ochoa, X., 2009. The Ariadne infrastructure for managing and storing metadata. IEEE Internet Comput. 13 (4), 18–25.

Wang, J., Robertson, S., Vries, A., Reinders, M., 2008. Probabilistic relevance ranking for collaborative filtering. Inf. Retr. 11 (6), 477–497.

Wiley, D.A., 2002. Connecting learning objects to instructional design theory: a definition, a metaphor, and a taxonomy. In: Wiley, D.A. (Ed.), The Instructional Use of Learning Objects. Online version (Retrieved June 26, 2013 from ⟨http://reusability.org/read/chapters/wiley.doc⟩). Agency for Instructional Technology, pp. 1–35.

Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, second ed. Morgan Kaufmann Publishers, San Francisco.

Xiang, R., Neville, J., Rogati, M., 2010. Modeling relationship strength in online social networks. In: Proceedings of WWW10, pp. 981–990.

Yu, Z., Zhou, X., Zhang, D., 2005. An adaptive in-vehicle multimedia recommender for group users. In: Proceedings of the Vehicular Technology Conference 5, pp. 2800–2804.

Zapata, A., Menéndez, V.H., Prieto, M.E., Romero, C., 2013. A framework for recommendation in learning object repositories: an example of application in civil engineering. Adv. Eng. Softw. 56, 1–14.