A Working Group Construction Mechanism Based on Text Mining and Collaborative Filtering

Kasthuri Arachchi, S.P.[1], Zhen-Rong Chen[1], Irugalbandara, T.C.[2] and Timothy K. Shih[1]

[1]Dept. of CSIE, National Central University, Taiwan.

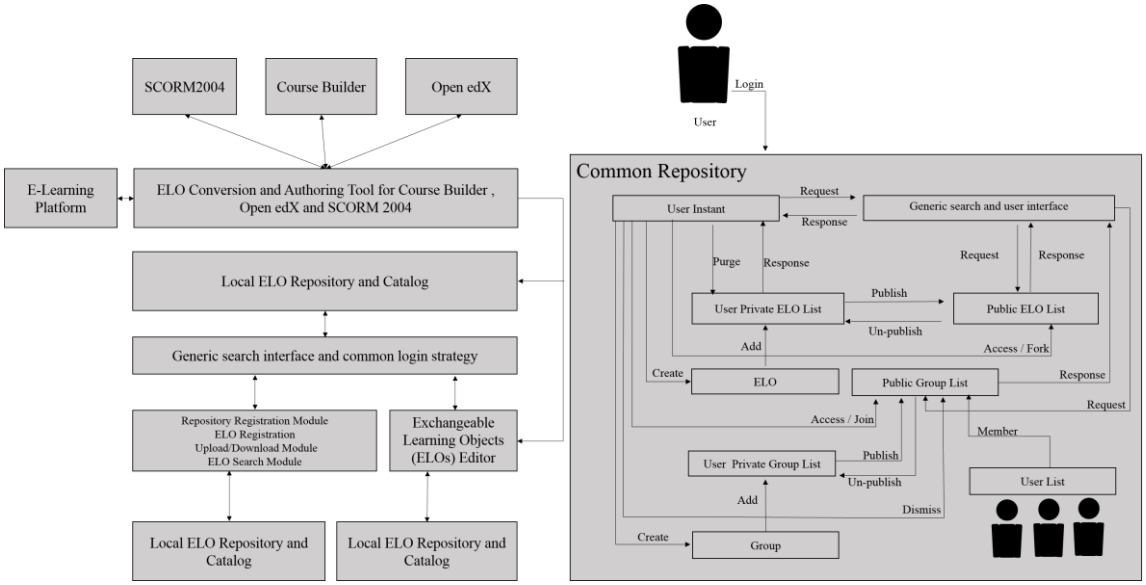
[2]Dept. of Physical Sciences, Rajarata University of Sri Lanka.

Email: {sandelik, timothykshih}@gmail.com

Abstract

Massive Open Online Courses (MOOCs) are popular in E-learning research domain with the advance of internet technology (Sa'don, Alias, and Ohshima 2014). MOOCs easily provide higher education courses for registered users as well as institutions or teachers who can offer courses in order to join more students than traditional education. However, producing high-quality learning materials may cause increase time, cost and efforts. For the purpose of reusing materials and reducing the cost of re-creating materials, the Learning Object (LO) concepts have been introduced. The content management systems which used these LOs are called Learning Objects Repository (LOR). The stored LOs in the repository can be easily searched by users. In this paper we introduce a working group construction mechanism for users on LOR. The proposed mechanism uses text mining technique to analyse the similarity of groups to construct prototypes of working groups. Then find the users' preferences about LOs based collaborative filtering to optimize constructed prototypes. Hence users on LOR can find quickly and easily their interesting learning materials via relevant working groups. This mechanism reduces the consuming time for re-creating learning materials by improving the quality of production.

This study is based on a Google MOOC FRA project (http://googleresearch.blogspot.tw/2015/03/announcing-google-mooc-focused-research.html). There are 3 parts of the system (Fig. 1 (a)) as: *conversion tool* between ELO (http://edxpdrlab.ncu.cc/), Course Builder, Open edX, and SCORM 2004; *Authoring Tool* for ELO; and *Repository for ELO* (Fig. 1 (b)). The user on the ELO repository can access the working groups which related to themselves and reduce the time consumed about re-creating learning materials and improving production quality.



(a) (b)

Fig. 1. (a) Overall scope of ELO project, (b) Overall system architecture of Common Repository

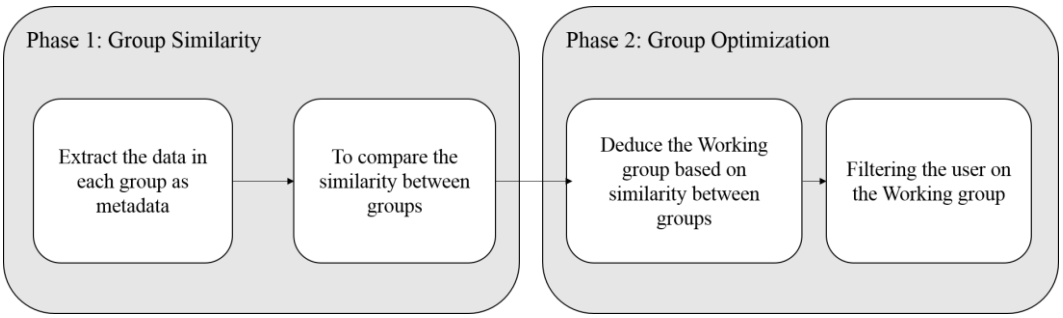


Fig. 3. Overall workflow of the proposed mechanism

The proposed working group construction mechanism for users on Common Repository (CR) can be divided into two phases as, *Group Similarity* and *Group Optimization* (Fig. 3).

Our system is developed with Python and HTML on Ubuntu. The Django framework is used on the server. We use Relational Database (RDB) to store the ELO contents. We postulate n users ($U = \{u_i | i = 1,2,3, \dots, n\}$), m ELOs ($L = \{l_i | i = 1,2,3, \dots, m\}$), and p groups ($G = \{g_i | i = 1,2,3, \dots, p\}$) on CR. Term frequency *tf* is,

$$tf_{i,j} = \frac{n_{i,j}}{\sum_{k=1}^{m} n_{i,j}}, \quad (1)$$

where, $n_{i,j}$ is the number of times the metadata M_i appears in a group j. For example, tf_{Taiwan} = 17/20 = 0.85 and $tf_{\text{Australia}}$ = 3/20 = 0.15 for a group with 20 people (17 Taiwan and 3 Australia) with a

nation metadata field. Term importance across all groups (Inverse document frequency *idf*) measures whether a term is common or rare across all groups.

$$idf_i = \log \frac{|G|}{|\{j:t_i \in gj\}|}, \tag{2}$$

where, $|G|$ is the total number of groups in the repository and $|\{j: t_i \in gj\}|$ is the number of groups that contain metadata M_i**.** For example, $idf_{\text{Taiwan}} = \log|2/2| = 0$ and $idf_{\text{Australia}} = \log|1/2| = 0.3$ for groups with 20 people in each, G1 (17 Taiwan and 3 Australia) and G2 (20 Taiwan). *tf-idf* measures which term is important enough to present a group as,

$$tfidf_{i,j} = \frac{n_{i,j}}{\sum_{k=1}^{m} n_{i,j}} * \log \frac{|G|}{|\{j:t_i \in gj\}|}. \tag{3}$$

The Jaccard Similarity Coefficient calculates similarity between groups as,

$$J(gi, gj) = \frac{|gi \cap gj|}{|gi|+|gj|-|gi \cap gj|}. \tag{4}$$

For example, Jaccard Index $j = 2/3 = 0.66$ for two groups G1 and G2, with n attributes (intersection and union of the attributes are 2 and 3). Pearson Correlation Coefficient (PCC) filters out the users that are not related to this working groups (Table 1) as,

$$\rho_{ui,uj} = \frac{E[(ui-\mu ui)(uj-\mu uj)]}{\sigma ui\ \sigma uj}. \tag{5}$$

For example, PCC $p = 0.54773$ for two users, U1 and U2, who have rated four ELOs with the rating history of U1 and U2 are {1, 2, 3, 4} and {2, 1, 1, 4}.

Table 1: The implication for the absolute value of PCC (Meijuan 2013)

The absolute value of PCC	**Relevance**
1	Perfect correlated
0.7~0.99	Highly correlated
0.4~0.69	Moderately correlated
0.1~0.39	Modestly correlated
0.01~0.09	Weakly correlated
0	Irrelevant

There were 30 simulated users, 50 simulated ELO courses, and four simulated groups on CR.

There were four original groups on CR, Group_1, Group_2, Group_3, and Group_4, with 5, 6, 5, and 6 members consecutively (Fig. 4).

ID	Name	Members
1	Group_1	Incents58 joshbruce851 Gire1986 Mott1961 Therstorted69 Thapt1936 Mined1976
2	Group_2	Pergersuse Timseat Knother Wheript Anorthems1964 Yese1993 Heiset Sheys1978
3	Group_3	Preacces Awassome Harys1960 Bacracks Hiontion83 Huble1943 Therteplied
4	Group_4	Experkee Tweeks Offard Squity67 Acep1952 Suraces Fuly1938 Arche1956

Fig. 4. The members of the original group

The proposed working group construction mechanism generated several working groups (Fig. 5).

11	Group_1_Group_4	Incents58 joshbruce851 Gire1986 Mott1961 Therstorted69 Experkee Tweeks Offard Squity67 Acep1952 Suraces
12	Group_2_Group_3	Pergersuse Timseat Knother Wheript Anorthems1964 Yese1993 Preacces Awassome Harys1960 Bacracks Hiontion83

Fig. 5. The members of the working group

The working groups, Group_1_Group_4 and Group_2_Gruop_3 has more like-minded persons than the original groups.

We proposed a working group deduction mechanism for users on CR. The proposed mechanism uses text mining technique to analyse the similarity of groups to deduce prototypes of working groups and find the users' preference about ELO based on collaborative filtering so that we can optimize these working group prototypes. For users on the LOR can easily discover the materials that they are interested via accessing the working groups which related to themselves and reduce the time consumed about re-creating learning materials and to improve production quality.