Challenges for Research into Open & Distance Learning:
Doing Things Better – Doing Better Things
Proceedings of the European Distance and E-Learning Network 2014 Research Workshop
Oxford, 27-28 October, 2014
ISBN 978-615-5511-00-4
ISSN: 2707-2819
doi: https://doi.org/10.38069/edenconf-2014-rw-0013

CLUSTERING BASED RECOMMENDATION OF PEDAGOGICAL RESOURCES

Brahim Batouche, Armelle Brun, Anne Boyer, University of Lorraine, France

Introduction

In France, seven DTUs (Digital Thematic Universities) allow open access to more than 24,000 OERs (Open Educational Resources). A DTU is a thematic repository of OERs, all validated by the academic community and indexed using SupLomFR (the French declaration for higher education of the LOM standard). The available pedagogical resources are of various nature (case study, lessons, exercises, simulation, virtual experimentation, additional materials to lessons, pedagogical kit, serious game, self-assessment, etc.) and various formats (pdf, audio, video, interactive or multimedia document, 3D, …). Each OER can be freely accessed from the DTU's portal, at any moment, by anybody, from everywhere. The main difficulty for a learner is to find the resources linked with his/her pedagogical objectives and his/her thematic background, when browsing the huge offer provided by a DTU. Facing the huge numbers of OERs and not familiar with the SupLOMfr indexing, most of users leave the DTU's portal without finding pertinent pedagogical materials. Thus, it is important to assist the user by a recommender system that suggests pertinent and adequate resources to him/her. In addition, it is more important to assist user in the context of open education.

The task of the recommender system could be viewed as the task of a librarian who helps users to find a pertinent book within a library. Books are classified by themes. A users request will be related to the theme which matches it the best. Our recommender system works on the same way: it classifies the UOH, (uoh.fr) the DTU dedicated to humanities, dataset in clusters (themes) in order to give recommendations according to the theme of interest of the user. If, because of the scarcity of resources in a specific theme, the user has already seen all the available resources, then the recommender will have nothing to suggest. In this scenario, the librarian's knowledge about close or linked themes allows to ensure nevertheless a high quality of recommendation. The question is how to do that automatically? The main difficulty of the task is that users are not registered, and we only can use the information collected during the current session for a given user.

This paper describes a recommender system relying on the last resources the user has consulted: the recommender system takes into account the fact that a resource has been accessed, as well as its description in SupLOMfr, if available. The interest of using information such as disciplines and keywords is to recommend the most adequate resources. Indeed, in the

context of e-learning, it is crucial to make accurate predictions: a recommender with a low quality of prediction is not acceptable. The quality of prediction can be highly affected by the scarcity of resources: a problem appears when the last resource viewed by a user is an isolated resource (no similar resource in terms of keywords and disciplines exists).

Knowledge about relationships between clusters can be automatically built by an unsupervised machine learning approach. Therefore, we use a clustering approach not only for its advantages as mentioned in Kim and Yang (2004) and Sarwar et al. (2002), but also to compute metadata about a dataset of resources, e.g. to build classes of resources and to determine links between classes. This knowledge presents a great advantage to solve the scarcity problem. It is the reason why we decide to recommend resources according to their description and also to their metadata.

Due to our applied characteristics and objectives, no method of clustering (Ghribi et al., 2010) is a better candidate for our use case than I2GNG (Improved Incremental Growing Neural Gas) (Hamza et al., 2008a; Hamza, 2008). The dynamic clustering of I2GNG and its capacity to build connection between classes, are the main advantages for our choice. I2GNG allows taking into account, in real time, the change and update done in the dataset. The results of I2GNG shows the distribution of resources and their scarcity, and the provided clusters are made up of a set of similar resources, which can be used by the recommender. Another advantage is that this model-based recommender system requires less processing time than one based on all resources descriptions, because the number of clusters is less than the number of resources. In summary, the use of I2GNG has fourfold objectives:

1. analysis scarcity of resources,
2. ensure the quality of prediction,
3. build knowledge-based recommender and
4. ensure scalable knowledge: clusters and connections between clusters.

The paper is organised as follows: Section 2 presents the related works. Our methodology is presented in Section 3. The results of the adjusted I2GNG algorithm are illustrated and discussed in Section 4, using a corpus provided by UOH. In Section 5, we conclude the paper and points work direction.

Related Works

As our approach focuses on accuracy prediction in e-learning context, by exploiting meta-data of resources, we present works focusing on predicting accurately, e-learning clustering and the existing approaches to index and exploit knowledge.

As showed by several works, such as Kim and Yang (2004), clustering of resources provides a higher quality of prediction, in studying the improvement of the quality of prediction, based on some attributes for each item. After comparing several neighbour selection methods, the authors conclude that the quality of the prediction of a recommender is improved by

ISBN 978-615-5511-00-4

clustering resources. The recommendation approach used in Sarwar et al. (2002) is based on a clustering of users. The authors propose an algorithm called ”clustered neighbourhood formation”, and the experimentations show that a recommendation system based on clustered neighbourhood has a higher quality of prediction because of the automatic improvement of clusters. Gribi et al. (2010) propose a clustering algorithm named “refined neighbour selection algorithm (RNSA)”. The algorithm uses the Pearson correlation coefficient between users, the transitivity of similarities and also the attributes of items. After experimentation and comparison of different methods of neighbours selection, the authors conclude that clustering-based recommender system using both refined neighbour selection and attributes can solve the large-scale problem of predicting accurately, without decreasing the prediction quality.

In the context of e-learning clustering, the clustering method is chosen according to the use case, such as available information. Kim and Yang (2005) cluster educational digital library in using LCA (Latent Class Analysis) (Xu et al., 2013), and shows that LCA provides better results than k-means algorithm (Magidson & Vermunt, 2004). LCA clustering is based on different types of parameters, but in our case, we have only the description of resources and no information about users. Lelu (1994) uses parallel affinity propagation (AP) to cluster large scale of e-learning resources. The experimentation shows that the clustering accuracy increases with the number of clusters. Unlike I2GNG, the statistical method AP does not allow to conserve meta-data about resources.

We distinguish two approaches of indexation:

1. Semantic description of resources and users profile, *e.g.*, through ontology models (Wang et al., 2008; Khribi et al., 2008). For pedagogical resources, the standard used to index the UOH resources is SupLOMFR (www.sup.lomfr.fr), which is based on an ontology. Coupling this ontology with a reasoning engine allows to index knowledge. The semantic reasoning becomes more complex when several factors (such as concepts and properties in the ontology) are taken into account.
2. Machine learning which refers to classification either unsupervised or supervised by experts (Manouselis et al., 2010). The use of machine learning (Golemati et al., 2007), as clustering methods, allows the extension of clustering-based recommendation, to knowledge- based recommendation.

The knowledge can be defined mathematically by building it automatically (based on a dataset), with I2GNG (Hamza et al., 2008a; Hamza et al., 2008b). The building process of the model by I2GNG, is based on available resources, and can update automatically the model, by creating gradually new classes, new connections, or removing existing classes and connections.

ISBN 978-615-5511-00-4

Methodology

Most of the time, the similarity between two resources is computed using the cosine similarity. Two resources are even closer that their similarity is high. As an illustrative example based on the UOH corpus, the nearest resource of the resource r_1 is r_2 according to the cosine similarity $(r_1, r_2) = 0.083$. The similarity between r_1 and r_2 is very low, *i.e.,* they are different resources, and we can verify it manually by opening their Web page of resources. We observe clearly that disciplines and keywords of these resources are different (French literature, transversal approaches). If we know *a priori* that the resource r_1 is isolated, then not to recommend is better than to recommend with a bad quality of prediction. This useful knowledge (r_1 is isolated), allows to avoid a bad quality of prediction.

To ensure the quality of prediction, we select only the non isolated resources. We use a clustering method to determine isolated resources. We have to remind that isolated resources could be considered as noise for most of the clustering algorithms. Then we decide to use a dynamic clustering method allowing creating and removing clusters during the learning phase, such as I2GNG. The I2GNG algorithm analyses data and then builds accordingly the neural network (in defining its structure and its weights), which will be exploited by the recommender and updated iteratively. The I2GNG recommendation consists in answering an *a priori* question; *is the target resource isolated?* yes/no. If yes, no recommendation will be given. Otherwise, nearest similar resources, which belong to the cluster provided by I2GNG, will be recommended.

Formalization of Pedagogical Resources

A clustering of the pedagogical resources of the UOH dataset will be done using the SupLOMfr indexing. The dataset, *ds*, is a set of resources $ds = \{\vec{r_1}, \ldots, \vec{r_i}\}$. After filtering noise in the description of a resource, by removing stop words, the pedagogical resource r_i is defined as a vector of maps between significant description words and their tf-idf (term frequency-inverse document frequency).

$$\vec{r_i} = [map_1, \ldots, map_k] \quad (1)$$

Where, map_i =(*word, tf-idf*), k the is variable size of resource $\vec{r_i}$, and *tf-idf* = *tf*idf*.

tf: the term frequency, $tf = \frac{|\{d_j : t_i \in d_j\}|}{|D|}$, where $|D|$ is the total number of words in the description d_j, and $|\{d_j : t_i \in d_j\}|$ the repetition number of word t_i in d_j. *Idf* is the inverse document frequency $idf = \frac{\log(|D|)}{|\{d_j : t_i \in d_j\}|}$.

Learning Function of Clustering

I2GNG is a neural approach. The learning process of a neural approach consists in defining a mathematical function *h*, which affiliates any resource from *ds* to its class c_j. Clustering

ISBN 978-615-5511-00-4

pedagogical resources consists in defining the cluster (presented as a neuron) c_j, the resources that belong to c_j, and the weight $n.\vec{w}$ of c_j.

A cluster (or neuron) contains several resources. The neuron weight represents the center of the cluster, *i.e.* it indicates the average description of all resources belonging to the cluster. Therefore, the vector size of a neuron weight evolves over time according to the learning set of resources. The size *m* of a neuron weight is greater than or equal to the size *n* of the vector coding a resource ($m \geq n$).

Let $\vec{r_a}$ be a resource and $n.\vec{w}$ be a neuron weight of the cluster of the resource, $\vec{r_a} \in R^n$, $n.\vec{w} \in R^m$, with $m \geq n$. The learning function is $h: R^n * R^m \rightarrow R$, where $h(\vec{r_a}, n.\vec{w})$ is the estimation of the membership relevance of the resource $\vec{r_a}$ to the cluster of the neuron *n*. Typically, $h(\vec{r_a}, n.\vec{w}) \in [0,1]$. The resources to recommend if the last viewed resources is r_a should be all the resources from the cluster with the neuron weight:

$$n.w_a = \arg\max_{n.\vec{w}} [h(\vec{r_a}, n.\vec{w})] \quad (2)$$

Adjusted I2GNG Algorithm

Neural approaches are computational models inspired by an animal's central nervous system (the brain) which is capable of machine learning as well as pattern recognition. The unsupervised neural I2GNG [6] is an incremental clustering method. The structure of the resulting neural network refers fully to the learning dataset ds, *i.e.,* the structure of the neural network will not be constrained by any initial condition. Clusters are created and removed dynamically, according to the learning dataset, without any degradation of the neural structure. The dynamicity of I2GNG allows a large tolerance to noise, such as isolated resources, because a cluster created from noise, will be detected implicitly. Despite these advantages, I2GNG requires to be adjusted to detect the isolated resources.

The input/output of I2GNG are respectively a set of resources *ds* (the UOH dataset) and a set of clusters (defined as a neural network). A cluster will be represented by a neuron, which at time of creation is considered as an embryo (*age=0*). For each added resource to the cluster, its age will be incremented, until attaining the mature age of the neuron. Each neuron n_j takes into account the characteristics of its resources, and is described by its weight ($n_j.\vec{w}$).

Let $\vec{r_i}$ be a vector representing a pedagogical resource, where $\vec{r_i} \in ds$. The winner neuron n_1 corresponds to the nearest neuron of the resource r_i, according to the cosine similarity (*cosineSim*), which is computed from the resource r_i and the set of neurons weights $n_j.\vec{w}$. We note n_m the neighbour neuron of n_1, *i.e.* a connection between n_1 and n_m exists.

In order to set the algorithm according to the context, we fix several parameters related to neural network elements (neuron and connection). The I2GNG parameters are: the mature age of the neuron n_j ($n_j.a_{mature}$), the max age of connection $c_i(c_i.a_{max})$, the adaptation

ISBN 978-615-5511-00-4

rate of a winner neuron (ϵ_b), the adaptation rate of winner neighbours (ϵ_m), and the neuron threshold $n_j.\partial$, which must be respected by resources belonging to n_j: $n_j.\partial = m_{n_j} + \propto \sigma_{n_j}$, where, m_{n_j} is the average similarity of resources that belong to the cluster of neuron n_j, σ_{n_j} is the standard deviation of similarities, and $\propto$ is a fixed parameter. We add a Boolean parameter n .*unitary*, which allows defining the isolated resources. If n .*unitary=true*, then the resource belonging to n will be isolated.

To build the neural network, the algorithm begins in checking if the input vector $\vec{r_i}$ does not match any weight of an existing neuron. If so, a new neuron will be created. When a neuron n_{new} is created, it will be affected to the embryo set of neurons, with an age equal to 0.

Initially, the neurons set are empty and, for each iteration, the winner neuron n_1 will be selected. If n_1 does not exist or if the similarity between n_1 and $\vec{r_i}$ does not exceed the threshold $n_1.\partial$, a new neuron will be created with $n_{new}.\vec{w} = \vec{r_i}$.

When the winner neuron n_1 does not satisfy the condition, we look for the second nearest neuron n_2. If n_2 *does* not exist or if the similarity between n_2 *and* $\vec{r_i}$ does not exceed the threshold $n_2.\partial$, a new neuron n_{new} and a connection between this neuron and n_1 will be created.

When the two neurons n_1 and n_2 satisfy the condition, a neuron n_{new} will not be created and the neuron weights of n_1 and n_m will be adapted, based respectively on ϵ_b and ϵ_m. Then, the age of the connection emanating from n_1 will be incremented in order to give more importance to n_1 versus its neighbours neurons. We add a connection between n_1 *and* n_2, if no. Otherwise, if a connection exists, the algorithm modifies the age of the connection to 0 to get importance to n_2 because it is the nearest cluster to $\vec{r_i}$.

A connection between two neurons means they are close. All connections with an age that exceeds a_{max} will be deleted, because this means that the departure neuron of connection has been solicited much more than his neighbouring neurons. Therefore we check the relevance of neighbouring neurons (destination neuron of connection). If the deletion of connection results in an isolation of a neuron $n_{isolated}$, this latter will be deleted if it satisfies the condition, where the threshold $[n_{isolated} + n_2].\partial = \beta * (n_{isolated}.\partial + n_2.\partial)$, and β is a fixed parameter. This threshold represents the acceptable similarity between the isolated neuron $n_{isolated}$ and its nearest neuron n_2. As the neural network evolves according to the new data, this condition allows avoiding degradation of the neural network structure. Then, the age of all neurons connected to n_1 will be incremented, and all embryo neurons whose age exceeds a_{mature} become mature neurons.

We adjust I2GNG algorithm because it builds (creates and removes) dynamically the clusters, in modifying the neuron weight $n.\vec{w}$, then the resources positioned at the border of the cluster may be similar to a resource belonging to the nearest cluster(s). This limit can generate a false

ISBN 978-615-5511-00-4

isolation of resources. Therefore, we validate the isolation of a resource, which is alone in its cluster. To do this, we compare the cosine similarity with its nearest resource $n_4.\vec{w}$ and the threshold $n.\partial$.

Experimentations and Results

We use the UOH dataset to test our algorithm. The UOH dataset is made of 1,294 resources. We use 70% of the dataset (910 resources) to learn our model and the remaining 30% (324 resources) will be used for the test. Our experimentation consists in analyzing the structure of the resulting neural network, in discussing the scarcity of resources and the evolution of the neural network between learning and validation steps.

Learning Step

The resulting neural network is decomposed into 152 connections and 583 neurones. The number of connections reveals that 152 clusters among the 583 (*i.e.* 26% of clusters) are close in term of similarity and can be merged in the future.

62.97% of treated dataset are divided into 131 clusters, where the average size is equal to 3.49 and the deviation is equal to 2.70. (37.03%) of treated dataset are alone in their cluster (isolated), *i.e.* there are 452 clusters having only one resource. It means the nearest neighbour of 37.03% of resources can be dissimilar.

Validation Step

30% of resources are used to test the resulting neural network of the learning step, where the resources will be affiliated to the nearest neurone. The resources belonging to the affiliated neurone will be presented to the user as a list of recommendations and sorted according to their similarity with the tested resources. We observe that 33.87% of OERs from the validation dataset are considered as new, *i.e.,* they do not match a cluster (it does not exist any similar resource from the treated dataset). 66.13% of validation dataset matches an existing clusters, *i.e.* they will be affiliated to their cluster. The resource(s) already into this cluster build the recommendations list.

Now, we observe the similarity *s*, between the last visited resource by a user and the first resource of the recommendations list. The similarities $s \in$ [0.57, 0.99], it means that 66.31% of resources will have good quality of prediction, and the rest has no recommendations list.

After validation step, 100% of UOH dataset is treated, and then the neural network was improved. To check the improvement of neural network and check if this improvement reduces the isolated resources, we observe the new structure of neural network. The number of connections and neurones becomes respectively 227 and 769.

The number of connections means that 39% of clusters can be merged in the future. However, 36.78 of the UOH dataset are alone in their cluster. It means that percentage of isolated

ISBN 978-615-5511-00-4

resources is reduced by 0.25 % comparing to the first step. And 63.22% of the rest are divided into 187 clusters, where the average size is equal to 3.80 and the deviation is equal to 3.24.

Conclusion

In an e-learning context, we have to ensure the capability of making accurate predictions, particularly in open education with anonymous users. The accurate prediction can be affected seriously by the sparsity of resources. To tackle this problem, we adapt the I2GNG algorithm to determine isolated resources and to discover knowledge about resources. As we exploit resources indexed with the standard SupLOMfr, our approach is adaptable to any pedagogical resources.

Based on our experimentation, 36.78% of UOH resources are isolated, this percentage corresponds to the probability of resources for which it is better not to recommend. The results of I2GNG give an important information for UOH management to improve their system, in showing isolated resources and their distribution.

Since our approach of recommendation detects the isolated resources, then we will have accurate predictions. We observe that 39 % of clusters can be merged in nearest future.

As a perspective, the proposed method of recommendation can be extended to consider:

1. the history of consulted resources of anonymous users (Bonnin, 2010), in taking into account: accuracy of prediction, speed of recommendation, and adaptability,
2. the group recommendation, which the case of Academic learning. This use case requires a hybrid of collaborative filtering techniques, and cluster-based recommendation.

ISBN 978-615-5511-00-4

References

1. Batouche, B.; Nicolas, D.; Ayed, H. and Khadraoui, D. (2013). *Using machine learning to recommend tourism activities for elderly people.* IEEE, EpsMsO.
2. Bonnin, G. (2010). *Vers des systmes de recommandation robustes pour la navigation web: inspiration de la modlisation statis- tique du langage.* PhD Université Nancy 2.
3. Ghribi, M.; Cuxac, P.; Lamirel, J.-C. and Lelu, A. (2010). Mesures de qualité de clustering de documents : Prise en compte de la distribution des mots clés. In *10ie`me Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances-EGC2010,* N.Be chet,Ed. Hammamet, Tunisia: Fatiha Sa¨ıs, Jan. 2010, p. 00.
4. Golemati, M.; Katifori, A.; Vassilakis, C.; Lepouras, G. and Halatsis, C. (2007). Creating an ontology for the user profile: Method and applications. In *1st RCIS, 2007.*
5. Hamza, H. (2008). *Application du raisonnement a partir de cas a l'analyse de documents administrates.* PhD University Nancy 2.
6. Hamza, H.; Belad, Y.; Belad, A. and Chaudhuri, B.B. (2008a). Incremental classification of invoice documents. In 19th - ICPR 2008. Tampa, United States: IEEE, 2008, (p. 4).
7. Hamza, H.; Belad, Y.; Belad, A. and Chaudhuri, B.B. (2008b). *An end-to-end administrative document analysis system.* IAPR International Workshop, 2008, (pp. 175–182).
8. Katakis, I.; Tsapatsoulis, N.; Triga, V.; Tziouvas, C. and Mendez, F. (2012). Clustering online poll data: Towards a voting assistance system. In *SMAP,* (pp. 54–59).
9. Khribi, M.J.; Kouthear, M. and Nasraoui, O. (2008). *Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval.*
10. Kim, T.-H. and Yang, S.-B. (2004). Using attributes to improve prediction quality in collaborative filtering. In *E-Commerce and Web Technologies.* Springer, (p. 1-10).
11. Kim, T.-H. and Yang, S.-B. (2005). An effective recommendation algorithm for clustering-based recommender systems. In *AI 2005: Advances in Artificial Intelligence*, (pp. 1150-1153). Springer.
12. Lelu, A. (1994). Clusters and factors: neural algorithms for a novel representation of huge and highly multidimensional data sets. In E. Diday, Y. Lechevallier, M. Schader, P. Bertrand & B. Burtschy (eds.), *New Approaches in Classification and Data Analysis, ser. Studies in Classification, Data Analysis, and Knowledge Organization,* (pp. 241–248). Springer Berlin Heidelberg.
13. Magidson, J. and Vermunt, J.K. (2004). *Latent Class Models.* (pp. 175–198).
14. Manouselis, N.; Vuorikari, R. and van Assche, F. (2010). Collaborative recommendation of e-learning resources: an experimental investigation. In *Journal of Computer Assisted Learning, 26(4),* (pp. 227–242).

ISBN 978-615-5511-00-4

15. Sarwar, B.M.; Karypis, G; Konstan, K. and Riedl, J. (2002). Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering. In *Proceedings of the fifth international conference on computer and information technology, vol. 1.* Citeseer, 2002.

16. Wang, W.; Zhang, H.; Wu, F. and Zhuang, Y. (2008). *Large scale of e-learning resources clustering with parallel affinity propagation.* ICHL.

17. Xu, B.; Recker, M.; Qi, X.; Flann, N.; Ye, L. (2013). Clustering educational digital library usage data: A comparison of latent class analysis and k-means algorithms. In *Journal of Educational Data Mining, 5(2),* (pp.38-68). http://www.educationaldatamining.org/JEDM/index.php/JEDM/article/view/21

Acknowledgment

This work has been performed under the PIA project Pericles (http://e-pericles.org).

ISBN 978-615-5511-00-4