

A *pseudo-supervised* approach to improve a recommender based on collaborative filtering ^{*}

José D. Martín-Guerrero

Digital Signal Processing Group, University of Valencia, Spain
jose.d.martin@uv.es

Abstract. This PhD Thesis develops an optimal recommender. First of all, users accessing to a Web site are clustered. If a user belongs to a cluster, the system offers services which are usually accessed by users from the same cluster in a collaborative filtering scheme.

A novel approach based on a users simulator and a dynamic recommendation system is proposed. The simulator is used to create the situations that one can find in a Web site. Introduction of dynamics in the recommender allows to change the clusters and in turn, the decisions which are taken. Since the system is based both on supervised and unsupervised learning whose borders are not too clear in our approach, we talk about a *pseudo-supervised learning*.

1 Description of the problem

The problem of finding an optimal recommender has two basic approaches: automatic customization approach and collaborative filtering [1]. In this work, we have chosen the second option because of the characteristics of our approach. Our strategy is firstly to cluster the users and afterwards to use the performed clustering for giving recommendations.

Since the Web sites can be formed by lots of links, it is important to reduce the dimensionality of the feature space where we cluster, from the number of services (a *service* is each one of the possible links which exist into a Web portal) to a lower number of labels (a *label* or *descriptor* is a set of services, all of them with similar contents).

The clustering algorithms must work by default in the space defined by labels because it is more feasible and the obtained clusters are more informative. These labels are usually chosen by the designer but sometimes no 'a priori' choice exists. Then, there are two possibilities: to cluster in the space defined by services or

^{*} This work has been partially supported by the Spanish Science and Technology Ministry project FIT-070000-2001-663, by the Valencian Culture, Education and Science Council project CTIDIA-2002-166 and by the University of Valencia project UV 01-15. I want to express my thanks to Drs. Emilio Soria-Olivas and Gustavo Camps-Valls for their direction and advices. I also want to express my thanks to Tissat, S.A., iSUM Department, <http://www.tissat.es/>, for its collaboration and technical support.

to obtain a good set of descriptors from the whole set of services. The latter can be achieved performing a Principal Component Analysis (PCA) [2] of the services which provides the linear combination of services (i.e., labels) with most information.

When desired clusters are not available, unsupervised learning algorithms must be used. In order to allow the evaluation of algorithms performance, we propose to simulate Web usage sites whose clusters are known. The aim is to develop a general Web simulator which could take into account the whole set of possibilities that could appear. Therefore, algorithms would be tested in every situation in order to know which algorithm is the best in each kind of Web site.

Afterwards, a recommender is implemented in order to give the best recommendation to users by using collaborative filtering. This recommender is based on the best clustering algorithm for this Web usage site, but not exclusively; in fact, a dynamic system which takes advantage from both supervised and unsupervised learning is developed. We call it, consequently, *pseudo-supervised learning*.

In Section 2, the proposed methodology is presented. In Section 3, it is described the developed work until now, its conclusions and planning for the future.

2 Methodology

The methodology begins with the development of the users simulator. We have taken into account two constraints which one can observe in all the real *log* files which were analyzed: the number of users who open up a new session decreases when the number of previously opened sessions increases; and, in each session, the number of users who access to a service decreases when the number of previously accessed services is higher. The constraints are obvious; when a user becomes an expert in a Web site, her navigation profile is straightforward but when she is a newbie, her profile is very random.

The simulator generates users in a label space. We have implemented several situations in our controlled experiments, covering most of those possibilities which can appear depending on the Web portal's characteristics. The simulator output offers information about the accessed services and labels by a user in a certain session. This information is coded into two tensorial matrices. The simulation of service accesses is achieved from label simulation. Once the label accesses are known, it is possible to find the number of services which are accessed because the relationship between labels and services is defined. Therefore, it is only necessary to take out the corresponding number of services from each descriptor.

The clustering works by default in the label space but it must be able to cluster users from information about the services because, sometimes, label information is not available. Anyway, it may be very difficult to cluster in the service space when its dimensionality is high. Moreover, one can take out more useful information from a label than from a service because the latter is too specific and it is very difficult to find a users' behaviour with only one service. In

fact, if predefined labels are not available, our proposal is to carry out a PCA of the services in order to reduce the dimensionality and to extract the best labels, because it will mean to cluster in the most informative space.

We have taken into account several clustering algorithms, such as, C-means, Fuzzy C-means, Expectation-Maximization algorithm, generic sequential and hierarchical clustering algorithms [3] and Self-Organizing Maps (SOM) [2]. Inclusion of ulterior algorithms in the system is trivial.

Once the clustering has finished, the recommender is based on collaborative filtering, i.e., those objects that were highly rated by other users with similar tastes are suggested. A user belongs to that cluster whose representative point (typically, its centroid) is nearest her. The clusters are dynamic since after the initial clustering, we can take advantage from the new users' behaviour by fitting the clusters to include this new information using a fuzzy scheme of Learning Vector Quantization (LVQ) [2]. This fitting implies that cluster representatives are adapted changing their position.

Summarizing, we propose a modular system with four parts. The first one is the simulator which allows to generate generic Web usage sites. The second is the choice of the best clustering algorithms; despite clustering is based on unsupervised learning, desired outputs are known because we generate them with the simulator. The third is the recommendation and the last one the fitting of the recommender including information from new users. We talk about *pseudo-supervised learning* to explain how the system works.

3 Work already done and tentative for the future

Users simulator and clustering algorithms comparison are finished. We have observed that very simple algorithms, such as C-means, work properly when the site is not complex. When the overlapping and the number of users' groups and labels increases, another approach must be used; we have observed that hierarchical algorithms work in medium complex sites but they fail when complexity is very high. The most complex sites need a more powerful tool: we have used a hierarchical SOM modifying the two-dimensional mapping with image digital processing and other variations.

Nowadays, we are working in the development of the adaptive system recommendation based on collaborative filtering and fuzzy LVQ. Future work will be related to non-linear tools applied to extraction of the best labels from the whole set of services.

References

1. Perkwitz, M., Etzioni, O.: Towards adaptive Web sites: Conceptual framework and case study. *Artificial Intelligence*, Vol. 118. Elsevier (2000) 245–275.
2. Ripley, B.D.: *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, UK (1996).
3. Theodoridis, S., Koutroumbas, K.: *Pattern Recognition*. Academic Press, a division of Harcourt Brace & Company, San Diego, CA, USA (1999).