Straight Through Grouping for Recommender System

Najma HAMZAOUI, Wafae BAIDA, Abdelfettah SEDQUI, Abdelouahid LYHYAOUI
Abdelmalek Essaâdi University, LTI Lab, ENSA of Tangier, Tangier, Morocco
PO Box: 1818, Tangier Principal

Abstract
In this paper, a new classification approach in recommender system is presented. First, we propose a co-classification of a data set to identify natural groups of similar objects. Then, we perform with a mechanism for the extraction process of these groups. The results are remarkable, especially for large scales data that requires more pre-treatment organizational evolutionary resources. This innovative method seems to give efficient and accurate grouping and recommendations.

Keywords: Collaborative Filtering, Group Technology, Recommender System, Co-Clustering

1. INTRODUCTION
With the huge amount of information flowing through the Web, it is increasingly difficult to find the information needed quickly and efficiently. However, with the advent of recommender systems (RS) during the 90 years [1, 2, 3], reducing information overload has become easy. Indeed, it is a system that aims to help users find interesting items, provide relevant information that meets their satisfaction and their real needs through a process of collecting, filtering and recommendation of the information. The RS are usually classified into three categories [4], content-based, collaborative filtering, and hybrid approach. Content-based methods are based on the characteristics of items to generate recommendations. On the other hand, CF is considered the technique of the most successful recommendation. Indeed, it is the most used in recommender systems for e-commerce. This technique allows recommending an item to a user based on the user profiles that are closest to him. In a recommendation system based on the CF data are presented in matrix, whose rows and columns are respectively based on a set of users and a set of items. Each user evaluates a set of elements by assigning certain values. The hybrid approach generates recommendations by combining the two approaches to content-based and collaborative filtering. To avoid the weaknesses of collaborative filtering techniques, research has focused on classification techniques based on models, in order to be more accurate and efficient. Thus, on the basis of user notes (ratings), these techniques include users (or items) forming clusters. This approach gives a new way to identify the neighborhood to make the recommendation, without using the entire database[2,3,7,8], it has produced several methods, like hierarchical clustering, [6, 9,10,11,12,13,14,15]. Another class of model, which is the evidence in collaborative filtering in recent years, is the matrix factorization [29, 30, 31, 32, 33, 34, 35]. Most of these methods can be grouped according to monofiltering and bifiltering approaches.

In this article, we adapt and apply the Bond Energy Algorithm (BEA) to perform a simultaneous co-clustering of the matrix. This allows us to find a natural clusters corresponding to the different components communities, and without a priori constraints on the number of class. Our work is structured as follows: Section 2 cites and summarizes the work related to recommender systems. In section 3, it describes the BEA algorithm, its application for co-classification and proposed for the extraction of classes representing natural communities’ solution. Finally, we conclude in Section 4 by the presentation of results and prospects.

2. RELATED WORK
The memory-based algorithms have shortcomings and are not suitable for a large system. For this, to achieve better prediction performance, researchers have proposed several approaches based on model-based CF [6]. The techniques of model-based, address better the problem of scalability in dealing with groups of examples, rather than the whole database. In a recommendation system, we find users who share the same tastes and interests. Consequently, we can combine to form a community. Many approaches address this problem in the literature provides several methods. Most of these methods can be grouped into two categories: monofiltering [17,18,19,20] or bifiltering (co-clustering). Despite the problem of scarcity, the biggest challenge is scalability FC. Many researchers have shown that the use of the technique of co-clustering is more robust to solve this problem, and it is a viable way to increase the scalability while maintaining a good quality of recommendation [17,19, 23]. The co-clustering involves both the grouping of users and items simultaneously. In [21], the algorithm used the simultaneous partitioning. The authors of [22, 23] used as a method of co-clustering, but introducing an analysis of the duality between users and items. They propose a system based on co-clustering and a new similarity measure algorithm. Thus, when the database is large, it is more appropriate to use the
ClustKnn method presented in [24, 25, 26]. The authors first compressed data by constructing a model of efficient clustering; the recommendations are generated using an effective approach based on the nearest neighbors. A summary of the FC-based clustering can be found in [5, 27]. A recent class of successful models of collaborative filtering is based on matrix factorization. Many methods have shown that the use of factorization methods for co-clustering gives a better result, as is the case of SVD, NMF, Tri-NMF, PMF, Non linear MFP MFP Bayesian methods, and NPCA [29, 30, 31, 32, 33, 34, 35].

3. The proposed approach
The size of networks requires the need to find methods can make them easy to manage. This requirement involves finding ways to structure it as groups with common characteristics. Nowadays, data using cross-classification or co-clustering comes from bioinformatics, text mining, but also the industry. In the industrial sector, the co-clustering is the name of the Group Technology (GT). It is a concept based on the identification and exploitation of the similarities and the similarity between the products and processes of design and manufacturing in order to rationalize production and reduce manufacturing costs [37]. In this sense, we present an algorithm named co-classification Bond Energy Algorithm (BEA), which is suitable for co-clustering of users and items. It is derived from the industrial world and generalizes co-classification algorithms working on GT. It is an algorithmic approach proposed by [28].

Methods of co-clustering process all rows and all columns in a table of data simultaneously seeking to obtain homogeneous blocks.

The data representation for the traditional collaborative filtering specially for item-based CF is based on the construction of an N×M item-user matrix U, showed in Table I. \( r_{ij} \) in the \( i^{th} \) row and \( j^{th} \) column of the matrix means the rating value for item \( i \) of user \( j \).

Table I. Binding energy of an element with its four nearest neighbors in the incidence matrix A

<table>
<thead>
<tr>
<th></th>
<th>( u_1 )</th>
<th>( u_2 )</th>
<th>( u_3 )</th>
<th>( u_4 )</th>
<th>( ... )</th>
<th>( u_M )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( v_1 )</td>
<td>( r_{12} )</td>
<td>( r_{13} )</td>
<td>( r_{14} )</td>
<td>( ... )</td>
<td>( r_{1M} )</td>
<td></td>
</tr>
<tr>
<td>( u_2 )</td>
<td>( r_{22} )</td>
<td>( r_{23} )</td>
<td>( r_{24} )</td>
<td>( ... )</td>
<td>( r_{2M} )</td>
<td></td>
</tr>
<tr>
<td>( u_3 )</td>
<td>( r_{32} )</td>
<td>( r_{33} )</td>
<td>( r_{34} )</td>
<td>( ... )</td>
<td>( r_{3M} )</td>
<td></td>
</tr>
<tr>
<td>( ... )</td>
<td>( ... )</td>
<td>( ... )</td>
<td>( ... )</td>
<td>( ... )</td>
<td>( ... )</td>
<td></td>
</tr>
<tr>
<td>( u_N )</td>
<td>( r_{N2} )</td>
<td>( r_{N3} )</td>
<td>( r_{N4} )</td>
<td>( ... )</td>
<td>( r_{NM} )</td>
<td></td>
</tr>
</tbody>
</table>

For a recommendation system, the use of TG, and specifically BEA results in the formation of communities. This algorithm is based on the rearrangement of rows and columns to reformatulate the matrix useful as a matrix of blocks. Indexes the rows and columns of these blocks represent the users and items of community members that are characterized by a strong similarity. Each block represents a strong association between users and items. Only a few methods have focused on complete extraction, specially, for large data, something that involves the search for a new method to perform an accurate and automatic extraction after the formation of natural blocks by BEA.

We can summarize the process of our Straight Through Grouping method (STGM) for Recommender System as follows:

- Set together those who are similar by applying BEA
- Detect, extract communities.

3.1 The propose of BEA
The purpose of BEA is to achieve a co-classification of a sparse matrix to identify groups of objects by making permutations of rows and columns of the incidence matrix. It also seeks to display and discover the associations and interrelationships between the groups with each other. It is based on the connection between an element of the incidence matrix A and the four nearest neighbors as illustrated in Table I. According to [28], these bonds can be considered as energy. Taking into account the frontier energy calculated, the permutation of rows and columns is made at the end to collect the elements of the matrix to create the group with maximum energy. The permutation is based on the value of coefficient of maximum energy (ME) as follows:

\[
\text{Energy}(r_{jk}) = r_{jk} \cdot (r_{jk} + r_{jk+1} + r_{jk-1} + r_{jk+1})
\]  (1)

Generally, the measure of effectiveness (ME) of a matrix A is the sum of its bond strength, where the bond strength between two nearest neighboring elements is defined as their product. The ME, is then given by:

\[
\text{ME}(A) = \sum_{i=1}^{M} \sum_{j=1}^{N} r_{ij}(r_{i+1,j} + r_{i-1,j} + r_{i,j+1} + r_{i,j-1})
\]  (2)
With convention $\mathbf{r} = r_{0-N-1} = r_{1-N} = r_{N-1} = 0$ and $\mathbf{A}$ is a non-negative matrix of dimension $N \times M$. ME maximization is taken over all $N! \times M!$ matrices, which can be obtained, from the input matrix by permutations of the rows and columns. This problem, as shown in [29], is reduced to two separate optimizations, one for rows and one for columns. Since the problems are equivalent, only the maximizing the sum of the bond lines or maximizing the sum of the bond columns should be discussed.

A algorithm, which exploits the nearest-neighbor feature, and is believed to be much faster and just as satisfactory (in the sense of achieving near optimal arrangements), has been developed and used successfully to determine array orderings corresponding to local optima of the ME. The algorithm is as follows.

**Table II. BEA ALGORITHM**

1. Place one of the columns arbitrarily. Set $i = 1$.
2. Try placing individually each of the remaining $N-i$ columns in each of the $i+1$ possible positions (to the left and right of the $i$ columns already placed), and compute each column's contribution to the ME. Place the column that gives the largest incremental contribution to the ME in its best location. Increment $i$ by 1 and repeat until $i = N$.
3. When all the columns have been placed, repeat the procedure on the rows. (The row placement is unnecessary; however, if the input array is symmetric, since the final row and column orderings will be identical.)

This algorithm has several important characteristics: It is finite and swift. The algorithm will always reduce an input array to pure block form and the final groupings and relations, however, have been found to be insensitive to the initial row (column) selected and their associated MEs have been found to be numerically close [29].

The algorithm will provide from the input matrix, a matrix of output as pure non-intersecting blocks or blocks checkerboard forms. In the case of the checkerboard form, blocks of non-zero matrix elements on the main diagonal represent pure groups, but the off-diagonal blocks indicate the relationships between the groups.

### 3.2 Extraction of communities:

BEA reforms the incidence matrix in the form of blocks with similar values, but this algorithm, don’t provide an automatic extraction of these blocks. Our proposed solution for the automatic detection blocks follows the following steps:

After applying the BEA incidence dimension matrix ($N \times M$), we propose to make a second reorganization based on the weight calculated $P(u_j)$ for each line (user) and $P(i_j)$ of each column (item).

$$P(u_j) = \sum_{l=1}^{N} R_{lj} \times w_l \quad \text{Avec} \quad w_l = 2^{M-l} \quad (3)$$

We consider that this weight is associated with each object. That it is the user or item. The basic building block of our method is an algorithm for partitioning that will be used on all users and all items which are represented by their weight. The usefulness of this weight appears when you decide to arranged them in descending order. Indeed, sorting of these weights of rows and columns transform the result matrix from BEA, in the diagonal form, show homogeneity of members of the same class (users or items) and will project the whole weight users on an axis and that of the items on another axis, so as to differentiate between the classes of users and items.

![Fig. 2a. Users sorted by the computed weights](image)
By applying the approach on a portion of the MovieLens data [38], we found the results shown in fig. 2.a and fig. 2.b. These results show the existence of groups of users / items with similar weight, hence belonging to the same community. For a user / item given, we find a remarkable variance between their weight, which explains the heterogeneity between the famous users / items. Therefore, these two elements can represent the boundaries of each community.

Then, the proposed approach tries to apply a solution to co-clustering matrix incidence of users and items, and search for partitions of users and items, based on the weight difference. Calculating the distance between the weight of users / items, is expressed by the difference between two successively ordered elements. Based on this, we can easily detect the boundaries of each class of users or items.

By this way, we just make a graphical presentation as shown in fig. 3.a and fig. 3.b for all distances calculated for users and for items. By observing the results presented, we can determine community members of users and items by choosing a threshold and projecting on the axis of the users and on the axis of items.

When we project the difference weight of the items on a dimension, classes of items are separated. For any pair of weight value of consecutive items separated by the greatest distance can be cut up and so on. We choose the best divisions. This method produces a set of clusters defined by intervals includes the items or users of one dimension.
This method is based on the output matrix of the BEA algorithm. It calculates a utility function for each line (items) and performs a descending sort on these values. The same procedure is applied to the columns (users) by calculating a utility and doing a sort on the values of this function. The values of each utility function (rows and columns) used to calculate the differences between successive values. This gives rise to two respective curves 3.3 and 3.4. It performs a data partition as the clusters do not overlap as shown in 3.3 for users and 3.4 for items. This method is quite effective, while these divisions are easily accepted because the fact of the important separation between clusters

4. CONCLUSION

This article attempts to address the need to improve the recommendation system using new techniques of clustering algorithm based on BEA, from Technology Group, used in the industrial field. It is very responsive and accurate for better co-clustering of large data systems. The proposed solution STGM is considered an extension of the algorithm BEA, especially in the extraction and determination of the number of classes by a transformation and a projection of the matrix resulting from BEA. Indeed, based on the blocks provided by the BEA algorithm, we used a measure function to find the classes of users and items. We thus obtained two corresponding projections for sorted users, and resources to determine classes. The method has been successfully tested on a portion of the database MovieLens. We are currently evaluating this method in comparison to others to apply it to various industrial problems in recommender system.

REFERENCES


[37] Walter Jean-Luc ; Mutel Bernard (Directeur de thèse) : Université de Mulhouse FRANCE, (Université de soutenance) » Study of group technology implementation. Application in the Holog international firm”