www.mbd.ase.ro/journal

Clustering of Respondents for Developing a Recommender and Targeting System: The Case of Assigning Education, Training and Employment Services

Klaus Bruno Schebesch
Vasile Goldis Western University Arad
kbschebesch@uvvg.ro
Radu Lucian Blaga
Vasile Goldis Western University Arad
buteniradu@yahoo.com

The context of recommendation of education and employment services, for example in the so called "on-demand job markets", provides various marketing implications. Our study is based on empirical data from respondents to the questionnaires, the issuers of demand for training and employment, mainly students, as well as on answers of employers from the western part of Romania.

We present an alternative analysis of the data without the use of socio-economic assumptions to guide results: using a simple evolutionary search enables explorative cluster formation. The comparison of clusters for student and employers opinions is based on the answer scores and illustrated by the emergence of distinct correlation structures for both types of respondents. At the end of our analysis we obtain the following information: the size of clusters, the comparison of cluster representatives, which turn out to be relative scores, which lead to cluster pairs with different degrees of opinion discrepancies between students and employers. High opinion discrepancies in cluster pairs inform which sub-population should be approached by educational recommendation while low discrepancies would conduct to job placement recommendations or services.

Key words: educational services, employment services, clustering, targeting, recommendation

JEL classification: M39, M54, J230.

1. Introduction

Suppliers of labor are in general organizations which may also provide educational services (schools, organizations for training and professional development). Private labor market matching services and different types of public employment offices share such activities. Generally, the dynamic nature of the socio-economic environment, especially that of labor markets determines the framework of the recommendation of such services. These services should include both group-based and also highly personalized variants. They should target adequate (often small enough) groups of applicants. The concept of "on-demand labor" also appears in the above mentioned context, being in part related to the organizational capacity of service suppliers to meet the actual needs of beneficiaries, subject to certain conditions on quality, structure and deadlines. However, on-demand labor may also be criticized and rejected by job-seekers on behalf of being viewed as a kind of "retrograde exploitation".

The educational and placement services, and the re-orientation of labor roles does not easily compare to any other category of services and markets. Relevant actions are often done using specific instruments (constrained by laws and regulations).

Within these services, the relationship between provider and customer, from the perspective of marketing, following Bradley (2001), is as follows: relationships are formalized, delivery is discontinuous; the provider-client contact has a high degree of customization, the demand fluctuates annually and the resilience of supply is low.

Educational services for the potential labor force, may be motivated by strong habitual, and cultural influences. Attitudinal aspects will probably play a major role in grouping persons for such services. All these features lead us to a series of marketing implications connected with targeting persons. In this sense the present research should be regarded as a specific and concrete computational mechanism as was proposed by Schebesch, Pop and Pelau (2010) in the context of "Computational Marketing". Specifically, we link certain aspects of explorative evolutionary search applied to grouping of persons by different recommending services in the labor markets of an economic region.

The resulting services will appeal to specific target groups, in being adaptable to many actual concerns like, e.g. on-demand labor job placement, by means of relatively continuous contacts with the clients, as observed in Vandermewe S. and Chadwick M.(1989) and Patterson P.G. and Cicic M.(1995) and also quoted by Bradley (2001).

The integrative vision presented here will be simplified and decomposed into algorithmic steps, starting from a set of empirical data (answers to questionnaires) offered by two distinct populations: 1) young potential employees and 2) employers from various activity areas of the analyzed region. In section 2 we describe the method of exploring clusters of students and clusters of employers. In section 3 we give some more details and context concerning our empirical data and its subsequent use. The findings of our computations will be described in section 4, proposing a way to target certain groups of persons by recommending placements on the labor market, and a complementary set of people, by recommending educational services. Finally, in section 5, we offer some conclusions and outlook.

2. Research methodology

The purpose of the study is to design a recommender and targeting mechanism that assigns services, tailored towards segments (clusters) of young people, by confronting them with an appropriate clustering of employers. These assignments can be exploited in the context of educational and employment placement services (on demand labor). The proposed analysis is based on the empirical data from respondents interviewed as part of research conducted by an European project entitled "A career for your life! Careers, advice and guidance."- SOP HR/161/2.1/G/132792. The project focuses on studying a regional labor market situation - involving a total of 1120 student respondents to questionnaires, respectively on a study concerning the barriers existing in searching and finding work. In the study participated a total number of 330 companies, with 96 of these specialized in technical activity domains, with 139 of companies working in the economic services, with 39 respondents coming from organizations active in the social and communities sector, with further 36 respondents coming from organizations working in health, and with 6 representing employers from the legal field and 7 respondents that do not fit in any of the above mentioned sectors.

The geographical region in which the two questionnaires were applied during the period June to November 2014, was Western Romania (especially the Arad and Timiș counties).

The paper uses clusterings as an exploratory tool without recurring biased pre-clustering of any kind (by using strictly coordinate-wise criteria, say). Instead k-means as a “simultaneous” grouping criterion is repeatedly used by applying simple evolutionary search. This search is used to restart k-means (which provides just local solutions) and to eventually select “improved” clustering by using a scoring criterion. For the merits and limitations of such restarts consult e.g. Bubeck et al. (2009). To implement the computations we used the powerful R-CRAN statistical programming platform.

The results of the questionnaires from our two populations (students and employers) are collected in two matrices, with line inputs denoting persons and columns stand for answers to a set of questions with a total of 78 answers. We limit our study here to the first 8 questions, the remaining information provided by questionnaires being discarded. Thus we use two restricted matrices: one for students or potential employees, with 1120 lines and 78 columns, and the other for employers, with 330 lines and 78 columns. The answers (column entries) are the same for both groups in number and content.

In the explorative search we use relative answer scores as the primary criterion for comparison between clusters. First we identify for both total populations those answers on which they differ strongly. This comparison already allows for certain coarse grained conclusions and recommendations. But these findings still need refinement.

To this end, both populations of respondents are clustered separately, repeating this procedure until we obtain relatively homogeneous groups of target persons. Unfortunately, for such tasks there exist many alternative clustering methods. For simplicity we limit ourselves to k-means, which more easily scales to even much bigger data. The usefulness of each clustering obtained during simple evolutionary search is measured by comparing relative answers scores for each single cluster (from both populations). Thus we compare the scores for each possible pair of clusters that belong to the two populations (students and employers). This however vastly increases the number of comparisons of such pairs if the number of clusters k is increased.

In the present study we limit ourselves to the exclusive use of 2-clusterizations ($k = 2$) during all epochs of the explorative search. This simple case also lends itself to a more natural interpretation of the findings. Hence, for pairs of clusters that are more similar in scores, we choose to recommend direct labor placement (leaving aside the technical matching details), and, for other pairs that are less similar, we would recommend educational services inspired by the discrepancies of the scores (for a stylized mechanism description see figure 1).

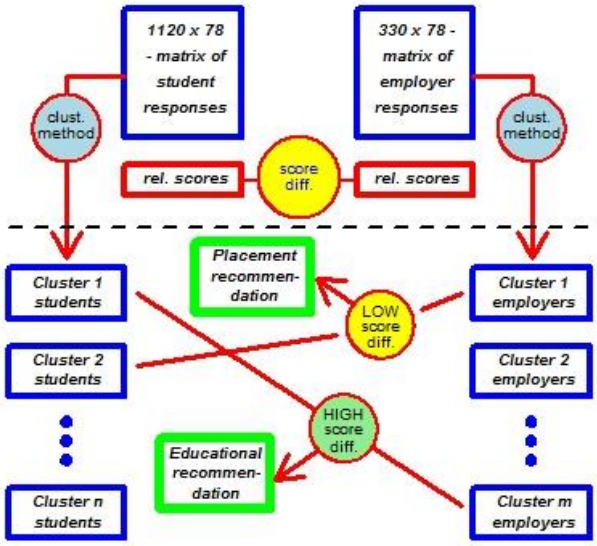

Figure 1. The evolutionary clustering algorithm scheme for our two analyzed populations, students (a part of the regional labor demand) and employers as a part of regional suppliers of labor.
Source: Own programming by utilizing R-platform for statistical software.

From figure 1 we also conclude that, in order to use the evolutionary clustering scheme to its full potential, one may encounter many resource consuming computational tasks, namely that of a) choosing one or more clustering methods, their concrete distance functions, of b) repeating the clustering step (here k-means with given $1 < k < N_i$, and with N_i the number of persons in population i) and of c) varying (a-)symmetrically the number of clusters per clustering k_s and k_e for the two populations. We will refrain here from diving into technical details of such a general evolutionary approach and we will stick to the simplest feasible case $k_s = k_e = 2$. Furthermore we restrict ourselves to evaluating the relative score differences.

3. Empirical Data and Scores

In our computations we use the „common answers" part of the questionnaire matrices, denoted Y_s and Y_e, respectively. They largely contain questions with „attitudinal" content, expressing opinions and atmosphere about the barriers to entry into the (regional) labor market. Both matrices have categorical entries which allow for simply summing up their columns (answers over persons) into „scores", by counting the number of given answers within each population. Such simple interventions are actually very useful, for instance, also for considering additional types of answers to more domain-oriented, technical questions, possibly with industry specific context, which would then imply using Rasch-type analysis expressing condions concerning „domain difficulty" and „average person score".

Our relative scores for the two populations are given by $s_s = (\mathbf{e}^T Y_s) / N_s$ and $s_e = (\mathbf{e}^T Y_e) / N_e$ respectively, with $\mathbf{e}^T$ being a vector of ones of appropriate length. By nature, these scores are all within the intervall of (0,1) and they are comparable across different population sizes.

4. Findings

In what follows we apply the procedure depicted in figure 1, further exploring the clusterings over a given number of epochs (1000, say), and restricting ourselves to one clustering method (*k*-means by using Euclidian distance). In figure 2 we start out by displaying the score differences found in the the two full populations of students and employers, respectively.

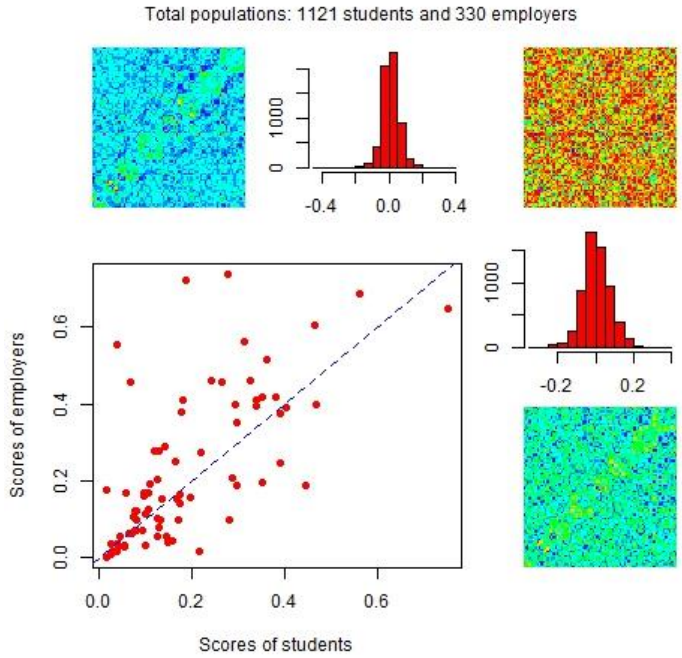

Figure 2. The relative score absolute differences between the answers of students and employers (main plot) complemented by the linear correlation of both populations (upper left inset for students and lower right for employers), the histogram of the correlation values without the main diagonal (mid-upper inset for students and mid-right for employers) and the absolute differences between the correlations of both population (matrix upper right inset).

Source: Own programming by utilizing R-platform for statistical software.

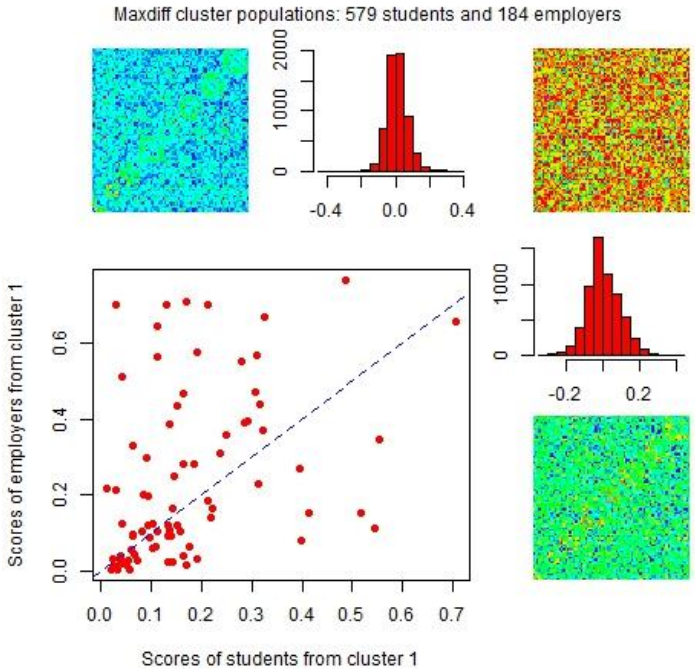

Figure 3. The same collection of plots as in figure 2, but for those clusters which imply recommendation of educational measures. This cluster pair (1, 1) is generated at epoch 589. For explanation see main text.

Source: Own programming by utilizing the R-platform for statistical software.

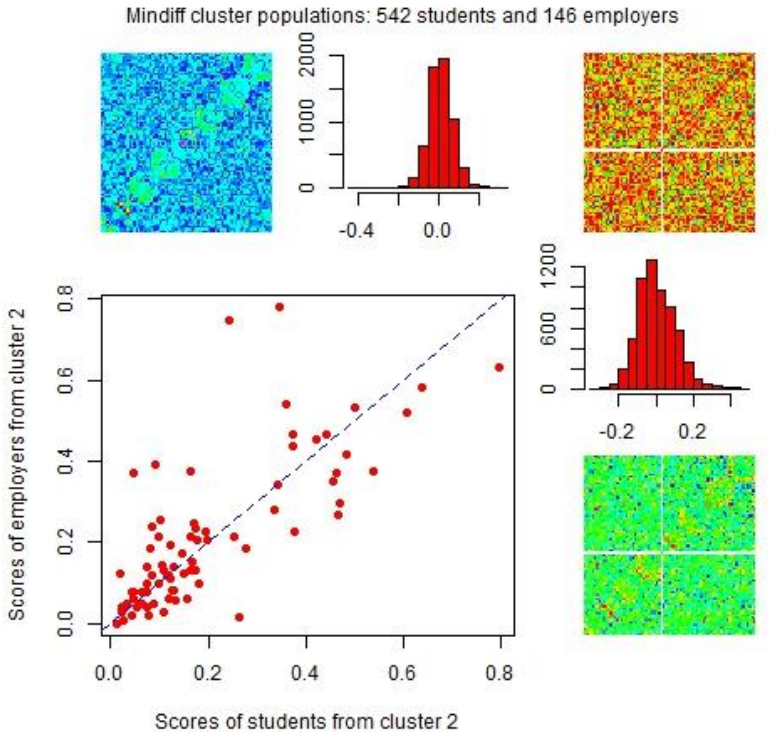

Figure 4. The same collection of plots as in figures 2 and 3 but here for those clusters which imply recommendation of job placements. This cluster complementary pair (2, 2) is generated at epoch 589. For explanation see main text.
Source: Own programming by utilizing the R-platform for statistical software.

From the figure we observe that confronting the score vectors ss and se (see previous sections) one obtains information about the *discrepancy* of scores (each point plots s_s(i) against s_e(i), for i = 1,2,...,78) as well as about the *distribution* of scores (i.e. here many small, few large ones). Note that there are slighly different correlation structures and distributions of correlation coefficients. Normal-like disributions indicate a proper (well colleted) data set of both populations (students and employers).

Figure 3 depicts a pair of clusters with larger discrepancies between the answers (views, opinions) of students and employers found at epoch 589, which implies seeking candidates for educational services from this cluster of students (cluster 1). Note also larger diffrences in correlation structure and distribution of correlation coefficients.

Figure 4 depicts the complementary pair of clusters with smaller discrepancies between the answers of students and employers found at epoch 589, which implies seeking candidates for job placement (matching) from this cluster of students (cluster 2) and firms (cluster 2). Note also larger differences in correlation structure and distribution of correlation coefficients. Correlation structure differs strongly from that of figure 3. One answer was not given by any of the 146 employers from this cluster (white crossing lines vizible in upper right hand side matrix).

Finally figure 5 depicts a summary of the simple evolutionary serch. The criterion for improvement (a kind of objective value) is maximizing the discrepancy between cluster pairs. Hence running the explorative serach for 1000 epochs produces improvements at increasingly larger distances in „time" (or epochs) as can be seen from the table in the left hand side inset.

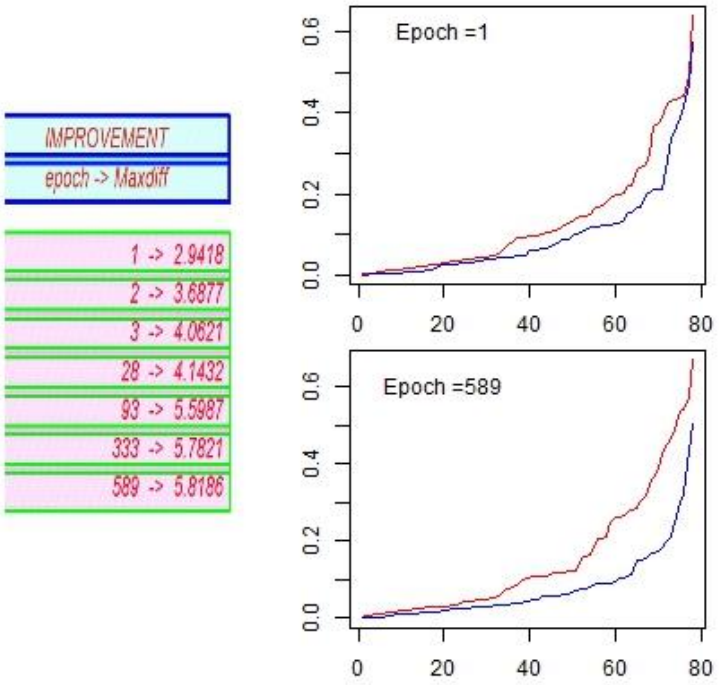

Figure 5. The simple stochastic search was conducted for 1000 epochs. Improvement in finding clusters with larger answer discrepancy is listed in the left column. The two right hand side inset plots depict the cluster pairs with largest and smallest discrepancy (upper and lower curves respectively). After the improvement at epoch 589 there was no more improvement in the run reported here.

Source: Own programming by utilizing R-platform for statistical software.

This explorative search shows that the clustering results are in fact quite stable in the sense of producing reasonably similarly sized clusters for population. This is confirmed by inspecting the entire history of the search (not shown here). As a last step in the present study we like to show what the more exterme discrepancies in the scores actually contain. To this end we first show the list of the common questions (shared by both populations) in table 1.

1	In your opinion the best situation is following
2	How much work-related experience did you accumulate during your studies? (e.g. volunteering, internship, part time / full time jobs)
3	Where are you looking for information on available jobs?
4	When you are searching for a job, what are the most useful informations ?
5	If you did apply for a job and you have been rejected, the reasons therefor were:
6	What do you think are the reasons for a graduate declining to work ?
7	What do you think are the reasons for a graduate to have difficulties in finding a job ?
8	What do you think are the most important barriers for young people wishing to find a job well suited to their aspirations and training levels ?

Table 1. The questions of the questionnaire common for students and employers numbered from 1-8. Each question many answer alternatives (in total 78, not shown here) and also allows for multiple answer.

Source: From the questionnaire developed by our research group and deployed during fall 2014 in the Western region of Romania

We refrain here from listing all the78 answers to these questions (the full list may be obtained from the authors). However, we are interested in displaying those question - answers pairs for which there are relative score differences of at least 0.4 (see also figure 5), that is the topics which generate the most disagreement between students and employers. It turns out that at epoch 589 there are 8 such top cases and for epoch 1 (not shown in the present material) there are 7 top cases. They largely conincide and we list

them together in table 2 below.

Orig. answ. code	Diff. in scores	In epoch 589	In epoch 1	Quest → answ	→ Explicit answer
[39]	0.670	1	1	5→ c	→ attitudine and presence not convincing
[37]	0.571	1	1	5→ a	→ fails to have competences required by employer
[58]	0.541	1	1	7→ c	→ candidate is pretending too high a salary
[46]	0.536	1	0	6→ b	→ candidate does not feel obliged to work
[27]	0.488	1	1	4→ c	→ candidate has (no) skills for the job
[38]	0.467	1	1	5→ b	→ defective communication
[50]	0.454	1	1	6→ f	→ low initiative / perseverence in job serach
[42]	0.429	1	0	5→ f	→ lack of experience
[8]	0.386	0	1	2→ a	→ lack of experience (a different question)

Table 2. The answers with the largest score discrepancy between students and employers, clusteringof epoch 598 (see also figure 5, the last improvement in serching for clusters with large) and the ansers also present in clustering epoch 1. The question – answers code is listed in column 5.

Source: Own programming by utilizing R-platform for statistical software

In table 2 (use table 1 to decode question) we observe a strongly - but not an exclusively - attitudinal context. Note the answers [46] and [42] do not appear in the top list of epoch 1 and [8] does not appear in the top list of epoch 598. The pro-eminence of question - answers pairs 6→ b and 5→ f give quite good reasons to design non-standard educational programs for young candidates in the regional labor market. Note also the persistence of the highly attitudinal question – answers pairs coded by the answers [39],[58], [50] and, in modern times, the somewhat surprising [38] which stands for "defective communication".

5. Conclusions and outlook

In this study we have developed, implemented and evaluated a simple explorative evolutionary search method for finding pairs of clusters from two functionally and socially distinct but interacting populations (students and employers from a region) which exhibit high or low discrepancies in relative opinion scores. In case of restricting ourselves to so called 2-clusterizations we find it also easy to interpret the different cluster pairs as candidate groups for educational services or for direct placement services, respectively. Here we did not consider fine grained matching mechanisms of persons (student – employer assignments) or concrete actions in order to support procedures for personalized "labor-on-demand".

However we may in principle deduce from the top score discrepancies list of evolved clusters new directions for concrete educational services which may have high attitudinal content (i.e. how to trigger attitude change in students, but maybe also in employers).

We propose that developing a more general procedure of multiple, heterogeneous clustering based on evolutionary principles, and perhaps not just on relative score differences, may be a worthwhile undertaking, both for labor placement companies and for labor-market oriented educational services organizations. This certainly will be the case when considering many more persons and much more detailed personal micro-data, e.g. concerning technological competences, risk propensities and private social positioning. Here again, appropriate clustering along the lines of Schebesch and Stecking (2014) may help to effectively hide sensitive personal data without losing much overall information about persons.

Acknowledgments

We thank the members of European project SOP HR/161/2.1/G/132792 lead by Horațiu Șoim of UVVG Arad and several other partner organizations from the Wesern region of Romania. Many thanks also go our students for cooperatively designing and entusiatically gathering the empitrical data.

References

BRADLEY, F. 2001. *Marketing Internațional*, Editura Teora, București

BUBECK, S., MEILĂ, M. & von Luxburg, U. 2009. How the initialization affects the stability of the k-means algorithm, 31 July 2009, accessed at http://arXiv:0907.5494v1

PATTERSON, P.G. & CICIC, M. 1995. A typology of services firms in international markets: an empirical

investigation. *Journal of International Marketing*, 3(4), 57-83

SCHEBESCH , K.B., POP, N. Al. & PELĂU, C. 2010. Marketing computaţional – o nouă paradigmă în marketingul contemporan, A New Paradigm in Contemporary Marketing – Computational Marketing, *Romanian Journal of Marketing*, 1, 36-73

SCHEBESCH , K.B., STECKING, R. 2014. Clustering for Data Privacy and Classification Tasks. In: D. Huisman, et al, (editors): Operations Research Proceedings 2013, Springer International, 397-402

VANDERMERWE, S. & CHADWICK, M. 1989. The internationalisation of services. *Services Industries Journal,* 9 (1), 79-93