Samsung!

Darya Lavrova, Ruslan Gaisin, Iskander Kareev, Rustem Salimov

Kazan Federal University

Abstract. Abstract

Keywords: keywords

1 Introduction

2 Demographic Clusterisation of the Gathered Data

The investigation was done on the basis of dataset consisting of user's age x_1 (column *webapi_agecateg*), gender x_2 (*gender*), marital status x_3 (*marital*), occupational status x_4 (*jposition*) and infomation on their Internet activity — the urls which they have visited.

To improve the robustness of the investigation and clearness of its results we have excluded the urls which were visited with less than 5 users. After the we had total of 526 user entries and 316000 entries on url visits.

Put U to be the set of all users, S to be the set of all sites (urls). By $S(A)$ denote the set of all sites which were visited by at least one user $u \in A \subseteq U$. By $U(s)$ denote the set of all users which visited the site $s \in S$.

2.1 Clusterization of Users by Demographic Attributes with Control on Diversification of Derived URLs Sets

On this part of the investigation we have recoded the values in the following way:

- for values of *marital*:
  "*Single*" → 0, "*In relations*" → 0.5, "*Married*" → 1;
- for values of *gender*:
  "*Male*" → 0, "*Female*" → 1;
- for values of *webapi_agecateg*:
  "*0..17*" → 1, "*18..24*" → 2, "*25..34*" → 3, "*35..44*" → 4, "*45+*" → 5;
- for values of *jposition*:
  "*employee*" → 1, "*executive*" → 1, "*jobless*" → 0, "*minor*" → 0, "*student*" → 0.5.

This allows easy application of classic clusterisation algorithms based on Euclid distance. Here we apply the hierarchic algorithm. The results of clusterisation are highly dependant on the scale of the variables. That is why all the variables were scaled by their means and variances. We bring in a vector of coefficients $\boldsymbol{w} = (w_1, w_2, w_3, w_4), \quad w_i \in [0, 1]$, so rescaled values are supplied to the clusterisation algorithm of the form:

$$(w_1x_1, w_2x_2, w_3x_3, w_4x_4).$$

Sites separation measure. Let us describe the considered way of choosing of coefficients $\boldsymbol{w}$ values.

Suppose that after the clusterisation with some $\boldsymbol{w}$ the users U are divided on k sets $C_1, C_2, \ldots, C_k$:

$$C_1 + C_2 + \ldots + C_k = U.$$

Let

$$r_{s,j} = \frac{\mathcal{N}(U(s) \cap C_j)}{\mathcal{N}(C_j)}.$$

By means of $r_{s,j}$ we define an intersection measure for the clusterisation $C_1, C_2, \ldots, C_k$:

$$M_I(\boldsymbol{w}) = \sum_s \Big(\sum_j r_{s,j} - \min_j \{r_{s,j}\} \Big).$$

For given number of clusters k we choose the weights $\boldsymbol{w}$ as the ones for which $I(\boldsymbol{w})$ is minimized.

Clusters number. Let $M_I(k) = \min_w M_I(k, \boldsymbol{w})$ be the intersection measure for clustering into k clusters. Fig. illustrates the dependance of the number of unique sites in each group and the value of M_I on value of k.

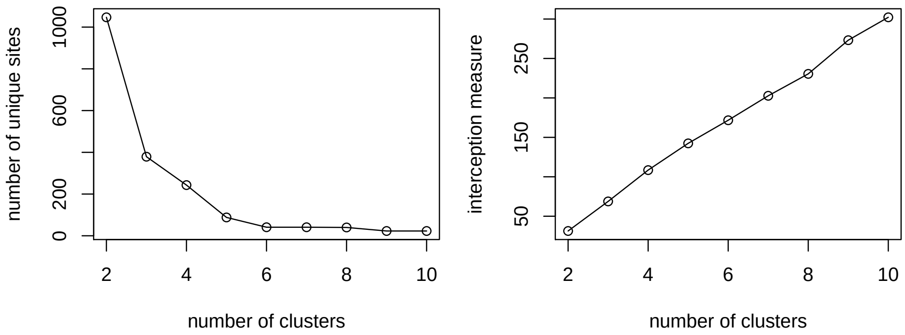

Fig. 1. tru-la-la

After the AIC algorithm results we chose $k = 6$ as the clusters number. For this the weights minimizing the interception measure M_I:

$$\boldsymbol{w} = (0.4, 0.4, 1.0, 0.4).$$

On that values of weights we might suggest that the age (to which correspond the weight 1.0) has the most has the most distinguishing effect on the visiting Internet sites.

		C_1	C_2	C_3	C_4	C_5	C_6
Number of users		71	282	119	33	9	12
marital	Single	14%	31%	22%	12%	0%	8%
	In relations	69%	25%	0%	61%	100%	58%
	Married	17%	44%	78%	27%	0%	33%
gender	Male	25%	38%	22%	79%	33%	67%
	Female	75%	62%	78%	21%	67%	33%
webapi_agecateg	0..17	0%	0%	98%	0%	100%	0%
	18..24	0%	64%	0%	36%	0%	0%
	25..34	0%	36%	0%	64%	0%	0%
	35..44	66%	0%	0%	0%	0%	58%
	45+	34%	0%	0%	0%	0%	42%
jposition	employee	58%	46%	3%	0%	67%	0%
	executive	42%	7%	3%	0%	11%	0%
	jobless	0%	0%	3%	94%	11%	75%
	minor	0%	0%	36%	6%	11%	25%
	student	0%	47%	55%	0%	0%	0%

Table 1. asd

The results.

Acknowledgments. The heading should be treated as a subsubsection heading and should not be assigned a number.

References

1. Smith, T.F., Waterman, M.S.: Identification of Common Molecular Subsequences. J. Mol. Biol. 147, 195–197 (1981)
2. May, P., Ehrlich, H.C., Steinke, T.: ZIB Structure Prediction Pipeline: Composing a Complex Biological Workflow through Web Services. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds.) Euro-Par 2006. LNCS, vol. 4128, pp. 1148–1158. Springer, Heidelberg (2006)
3. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)

4. Czajkowski, K., Fitzgerald, S., Foster, I., Kesselman, C.: Grid Information Services for Distributed Resource Sharing. In: 10th IEEE International Symposium on High Performance Distributed Computing, pp. 181–184. IEEE Press, New York (2001)
5. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The Physiology of the Grid: an Open Grid Services Architecture for Distributed Systems Integration. Technical report, Global Grid Forum (2002)
6. National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov