Samsung!

Darya Lavrova, Ruslan Gaisin, Iskander Kareev, Rustem Salimov

Kazan Federal University

Abstract. Abstract

Keywords: keywords

1 Introduction

2 Clusterisation of users by demographic attributes

The goal was to provide the clusterisation of the set of users on several groups based on their demographic attributes (like age, occupational status and so on) and the sites they visited. The expectation was to get an interpretable insight on the connection between users' social status and their preferences on Internet sites.

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.

Notion and attributes recoding. Let U be the set of all users in the dataset and S be the set of all sites (urls). By $S(u)$ denote the set of all sites which were visited by user $u \in U$. By $S(A)$ denote the set of all sites which were visited by at least one user from the set $A \subset U$. By $U(s)$ denote the set of all users which visited site $s \in S$.

The investigation was done using the following demographic attributes: marital status (a_1), gender (a_2), age category (denoted as a_3) and occupational status (a_4). So to each $u \in U$ a vector of attributes $\boldsymbol{a} = (a_1, a_2, a_3, a_4)$ is assigned.

For simplicity of application of common clusterisation algorithms and to preserve the natural order of the values, all of the demographic attributes were recoded from categorical to numerical format in the following way:

- for a_1 (*marital* column in original dataset):
  "*Single*" → 0, "*In relations*" → 0.5, "*Married*" → 1;
- for a_2 (*gender* column in original dataset):
  "*Male*" → 0, "*Female*" → 1;
- for a_3 (*webapi_agecateg* column in original dataset):
  "*0..17*" → 1, "*18..24*" → 2, "*25..34*" → 3, "*35..44*" → 4, "*45+*" → 5;

- for a_4 (*jposition* column in original dataset):
  "*employee*" → 1, "*executive*" → 1, "*jobless*" → 0,
  "*minor*" → 0, "*student*" → 0.5.

The clusterisation of U is done by the values of attributes (a_1, a_2, a_3, a_4) using complete linkage hierarchical clustering with Euclidean distance. Let us note that the clusterisation were done using *hclust* method of *R* statistical package. Since those algorithm are sensitive to the scale of the data, the attributes a_1, a_2, a_3, a_4 are rescaled to have zero mean and unit variance throughtout the dataset. We denote the values of the attributes after recoding and rescaling by $\boldsymbol{x} = (x_1, x_2, x_3, x_4)$.

Due to aforementioned sensitivity of the clusterisation algorithm to the scale of input variables, we can regulate the importance of each of the attributes by bringing in weights $\boldsymbol{w} = (w_1, w_2, w_3, w_4)$ where $w_i \in [0, 1]$, $i = 1, 2, 3, 4$. Thus the clusterisation is done on the values of vector $(w_1x_1, w_2x_2, w_3x_3, w_4x_4$.

Let k be the number of clusters on which U supposed to be clusterised. By $\mathfrak{C}(\boldsymbol{w}, k)$ denote the resulting clusterisation for given k and weights $\boldsymbol{w}$, so

$$\mathfrak{C}(\boldsymbol{w}, k) = \{C_1, C_2, \ldots, C_k\} : \quad C_1 \cup C_2 \cup \ldots \cup C_k = U.$$

Let $S_i = S(C_i)$ — the image of users clustarisation $\mathfrak{C}(\boldsymbol{w}, k)$ to the set of all sites S. We want to pick values of weights in such way that the sets $S_1, ..., S_k$ are well separated and there is no much intersection between them, i.e. our clusterisation results in users clusters which are well distinguishable in terms of the sites users visit. In the following our approach to that task is described.

Weights and sites intersection measure. Suppose the number of clusters k is fixed. Let us describe the algorithm for choosing weights $\boldsymbol{w}$. Let

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

where $s \in S$, $C_j \in \mathfrak{C}(\boldsymbol{w}, k)$ $(1 \leq j \leq k)$, and $\mathcal{N}$ is the number of elements in its argument.

By means of $r_{s,j}$ we define an intersection measure for the clusterisation $\mathfrak{C}(\boldsymbol{w}, k)$:

$$M_I = \sum_{s \in S} \Big(\sum_{j=1}^{k} r_{s,j} - \min_{j: 1 \leq j \leq k} r_{s,j} \Big).$$

Measure M_I adds a "penalty" for sites which are visited by users from several different clusters.

We choose the weights $\boldsymbol{w}$ as the ones for which M_I is minimised. We solve this minimisation problem computationally with the precision of 0.2 points.

Clusters number k. In our investigation we studied the influence of clusters number k to two parameters: the number of "unique" sites (i.e. the ones which occur only in one of the sets $S_j, \quad 1 \leq j \leq k$) and the value of the interception measure M_I (for "optimal" weights $\boldsymbol{w}$). Those dependancies are depicted in Fig. 1.

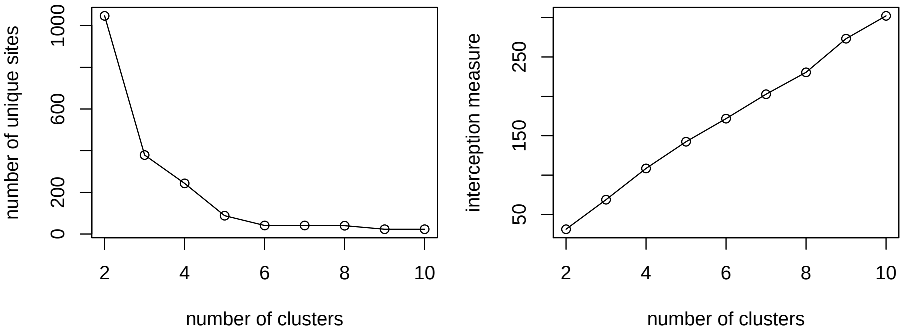


Fig. 1. The number-of-unique-sites vs k plot (on the left) and M_I vs k plot (on the right)

Taking into account the aforementioned parameters and ffter the AIC algorithm results we chose $k = 6$ as the clusters number. For that k the weights minimising the interception measure M_I are found to be:

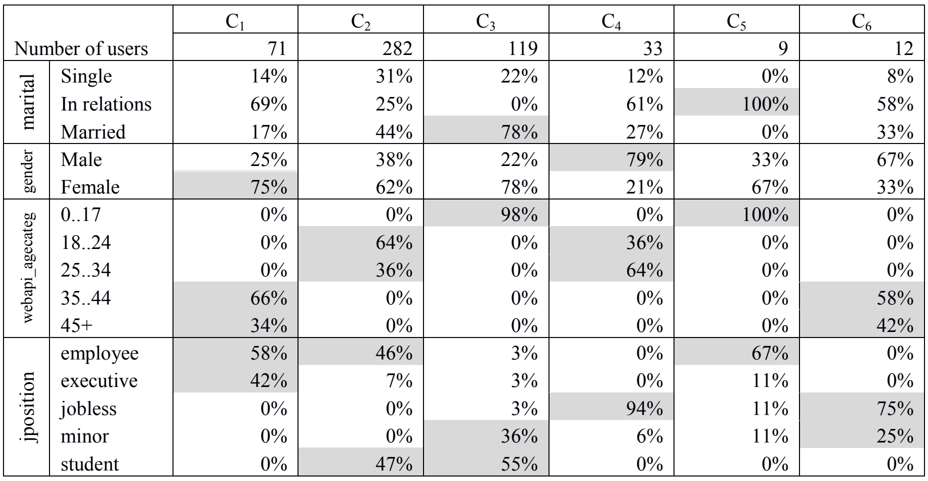
$$\boldsymbol{w} = (0.4, 0.4, 1.0, 0.4).$$

Let us recall that the weight w_3 corresponds to the age of the user. Due to that we might suggest that the agehas the most the most distinguishing effect on the preference of the users for the Internet sites.

The results and their interpretation. Some information on the resulting clusters are provided in Tables 1 and 2. The Table 1 contains the demographic composition of the clusters. The Table 2 contains 10 subject categories of sites (the mapping of sites to the categories were done according to Yandex database) most specific to the relevant clusters.

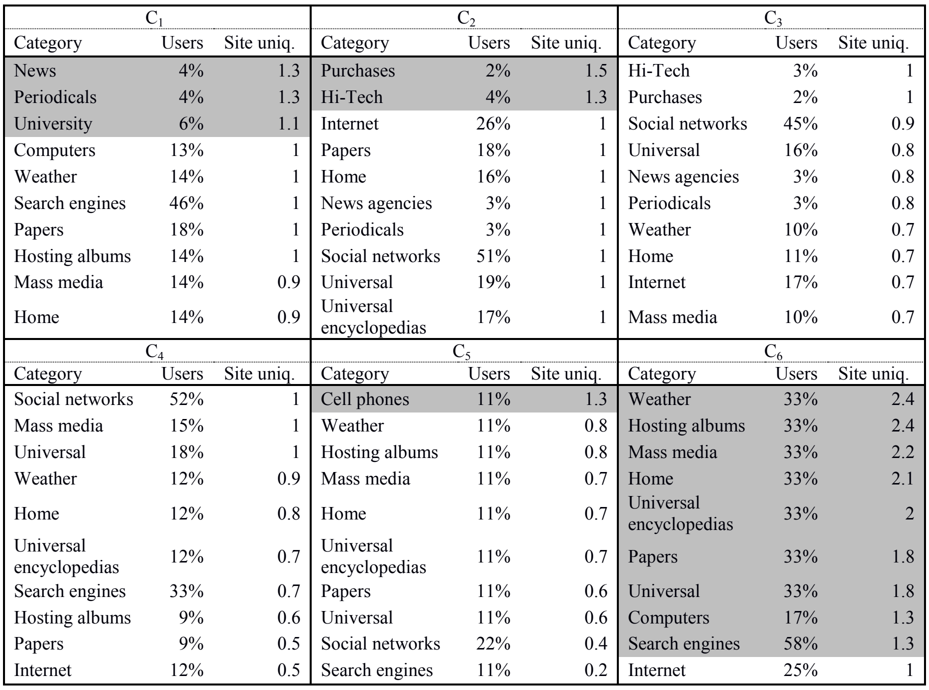
3 Dependency analysis between demographic attributes of a users and sites they visited

One of the problems was the task of checking the independence of the attributes of the data and sites categories. For this purpose chi-squared test is usually used if explored indications are qualitative (see [1]). So we had to build a contingency tables and to make decisions about hypothesis of independence.



		C1	C2	C3	C4	C5	C6
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. The demographic composition of the clusters. The most distinctive attribute values for the clusters are greyed out.



C1			C2			C3		
Category	Users	Site uniq.	Category	Users	Site uniq.	Category	Users	Site uniq.
News	4%	1.3	Purchases	2%	1.5	Hi-Tech	3%	1
Periodicals	4%	1.3	Hi-Tech	4%	1.3	Purchases	2%	1
University	6%	1.1	Internet	26%	1	Social networks	45%	0.9
Computers	13%	1	Papers	18%	1	Universal	16%	0.8
Weather	14%	1	Home	16%	1	News agencies	3%	0.8
Search engines	46%	1	News agencies	3%	1	Periodicals	3%	0.8
Papers	18%	1	Periodicals	3%	1	Weather	10%	0.7
Hosting albums	14%	1	Social networks	51%	1	Home	11%	0.7
Mass media	14%	0.9	Universal	19%	1	Internet	17%	0.7
Home	14%	0.9	Universal encyclopedias	17%	1	Mass media	10%	0.7
C4			C5			C6		
Category	Users	Site uniq.	Category	Users	Site uniq.	Category	Users	Site uniq.
Social networks	52%	1	Cell phones	11%	1.3	Weather	33%	2.4
Mass media	15%	1	Weather	11%	0.8	Hosting albums	33%	2.4
Universal	18%	1	Hosting albums	11%	0.8	Mass media	33%	2.2
Weather	12%	0.9	Mass media	11%	0.7	Home	33%	2.1
Home	12%	0.8	Home	11%	0.7	Universal encyclopedias	33%	2
Universal encyclopedias	12%	0.7	Universal encyclopedias	11%	0.7	Papers	33%	1.8
Search engines	33%	0.7	Papers	11%	0.6	Universal	33%	1.8
Hosting albums	9%	0.6	Universal	11%	0.6	Computers	17%	1.3
Papers	9%	0.5	Social networks	22%	0.4	Search engines	58%	1.3
Internet	12%	0.5	Search engines	11%	0.2	Internet	25%	1

Table 2. Some of the categories of sites which are visited by the users of the relevant clusters. The column "Users" shows the proportion of users in the cluster which have visited the sites of the category. The column "Site uniq." shows how mush the proportion of users for the category within this cluster is more than in any other cluster.

Originally it was planned to analyze the relationships between the features of respondents and visits to the sites. However, for most sites this version of the analysis is not suitable due to the fact that they have a small number of visits; the same time for the sites with big number of visits(vk.com, google.com, and so on) to allocate a significant relationship is not possible. Therefore, it was decided to group the original sites into categories for further analyzing. The categories with large number of visits (more than 1000) were selected for the analysis, because for the categories with fewer visits the study is not applicable due to the limitations of test. For each user and each category has been allocated fact of the visit (variable that possesses values 0 or 1). After that contingency tables were built and by these tables the hypotheses of independence were tested with a significance level 0.05.

It should be noted that two most popular categories: "Social networks", "Bots" - in which the number of visits by much more than in all other categories, depending on the features of the respondents users could not be found. That is a logical result, because websites of the most popular categories should be visited by users regardless of age, sex and other features.

A little bit strange result is that a visit to any one of the categories doesnt depend on the gender of the respondents. Perhaps this is due to a not enough large sample size.

Some interesting relations:

- Between the **age** attribute and the category "**Newspapers**" (with a high significance level **0.0015**) The result shows the benefit of young people aged 0 to 24 years of age and older people 45+ are much more likely not visit sites category "Newspapers".
- Between the **marital status** and the category "**Dating**" (the significance level of **0.026**). As expected, unmarried single people visit sites "dating" category.
- Between the **age** attribute and category of websites "**Dating**" (with a high level of significance **0.009**). Older people visit dating sites are more likely than younger ones.
- Between the **age** attribute and the "Internet" category (with a high level of significance **0.01**). It confirms the assumption that young people are more likely to visit sites with such a category.
- Between the **age** attribute and category of "**Information Agency**" (with a significance level of **0.022**). Here we see an unclear relation: young people under the age of 17 years, significantly more likely to visit sites with the category News agencies.
- Between marital status and the category "**Shopping**" (a high level of importance of **0.0053**). Married people, or irrelevant, significantly more likely to visit sites category Shopping, in turn, single people opposite.
- Between the **age** attribute and the category of websites "Work" (the significance level of **0.013**). Young people were significantly more interested in work sites category.

- Between the **age** attribute and the category "**Universal Encyclopedia**" (high significance level **0.0023**). Young people between 18 and 24 years old visit sites og the category "Universal Encyclopedia" rare than younger participants (0..17) and slightly older participants (24..34).
- Between the **workplace** and the category "**Universal Encyclopedia**" (high significance level **0.0082**). For this feature the obvious relation that students use the sites of this category more often than other groups figured out again. Complete workers also use sites of this category.
- Between the **age** attribute and the category "**Humor**" (significance level **0.04**). For this category we get the obvious connection that young people aged 18 to 34 years visited sites category humor more often.
- Between the **labor** attribute and category of websites "**Humor**" (significance level **0.03**). The connection that students visit sites in this category more often is highlighted here.

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

References

1. Wikipedia: Chi-Square Test https://en.wikipedia.org/wiki/Chi-squared_test