## § 2.2. Анализ и предварительная обработка данных анкет руководителей ИТ-предприятий г. Ульяновска

Коллективом ученых УлГТУ (Н.Г. Ярушкина, Т.В. Афанасьева, О.В. Шиняева, К.В. Святов, Л.М. Валкин, Д.А. Ефремов, К.Г. Калимуллин и др.) в рамках проекта «Исследование ИТ-кластера Ульяновской области» в 2013г. проведено анкетирование руководителей ИТ-предприятий г. Ульяновска. Собраны данные по 87 предприятиям по 39 вопросам анкеты (см. [22 с.112 ]).

При анализе и обработке анкетных данных возникли следующие проблемы. Во-первых, кластеризация требует объединения близких по смысловому значению параметров в группы и сравнения объектов (ИТ-предприятий) по этим группам параметров (см. § 2.1).

Во-вторых, анкетные данные разнородны. Среди них числа (рубли, количество сотрудников и т.д.), проценты, термины. Требуется перевод лингвистических переменных в числовые, требуется приведение всех данных к одному типу единиц измерения.

В-третьих, велик разброс данных в пределах одной группы. Теория кластерного анализа показывает, что в этом случае кластеризация затруднительна. Требуется нормировка данных.

В-четвертых, часть вопросов анкеты руководители предприятий оставили без ответа. Предприятия имеют разный уровень развития и данные некоторых предприятий составляют долю, близкую к 0, по сравнению с лидерами. Это создает трудности при обработке данных. Матрица 87 предприятий на 39 ответов имеет значительные пустующие поля. Требуется алгоритм уменьшения размерности матрицы, исключения из рассмотрения предприятий и вопросов анкеты с нулевыми или малыми значениями параметров.

Предварительная обработка данных в нашем случае включает в себя:

– приведение данных к однородному виду:

– ( Uniformity);

– анализ;

– редукцию (редукция в технике – сокращение, уменьшение силы движения, напряжения): – ( Reduction).

Предварительная обработка данных производилась в указанном порядке: .

Входные данные программы можно представить в виде таблицы 1. По горизонтали приведена нумерация параметров согласно списку из § 2.1.

**Таблица 1.**

**Результаты анкетирования руководителей ИТ-предприятий Ульяновской области за 2012г. (87 предприятий, 39 вопросов)**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Группы** | **Область деятельности компании** | **Финансово-экономические показатели** | | | | | | | |
|  | **Название компании** | **1.1** | **2.1 (руб)** | **2.2(руб)** | **2.3** | **2.4** | **2.5** | **2.6** | **2.7** | **2.8** |
| 1 | ИнтелСофт | Разработка собственных программных продуктов, Обучение IT-технологиям | 0 - 500 тыс | 0 |  |  | 0 | 100 | 0 | 0 |
| 2 | СимбирСофт |  |  |  | 50 |  | 100 |  |  |
| 3 | ИП Хомяченко | 0 - 500 тыс. |  |  |  |  | 100 |  |  |
| 4 | JackNyfe (Echo) |  |  |  |  |  |  |  |  |
| 5 | Группа Компаний ИТМ | 10 - 50 млн. | 0 - 500 тыс. |  |  |  |  |  |  |
| 6 | ООО "Купи батон" | 1 - 10 млн. | 1 - 10 млн. | 0,7 | 0,8 | 0 | 100 | 0 | 0 |
| 7 | JackNyfe(Echo) | 100 млн. | 50 - 100 млн. |  | 100 | 0 | 100 | 0 | 0 |
| 8 | ООО АИСТ | 1000 тыс |  |  |  | 0 | 100 | 0 | 0 |
| 9 | ООО"Телеком.ру" |  |  |  |  |  |  |  |  |
| 10 | ITmaster | 1000 тыс |  | 20 |  |  | 100 |  |  |
| 11 | Effective Script |  |  |  |  |  |  |  |  |
| 12 | Веб-студия Creater | 1 - 10 млн. | 1 - 10 млн. | 50 |  | 0 | 100 | 0 | 0 |
| 13 | ооо "андер девелопмент" |  |  |  |  |  |  |  |  |
| 14 | ITECH.group | - 500 млн. |  | 40 |  |  |  |  |  |
| 15 | ООО "АктивУчет" | 0 - 500 тыс. |  | - 7 % |  |  |  |  |  |
| 16 | ООО «Решение» | 1 - 10 млн. | 0 - 500 тыс. | 5 | 1 |  | 100 |  |  |
| 17 | ООО «Тауруна» | 1 - 10 млн. |  |  |  |  | 100 |  |  |
| 18 | ООО "УмКо" | 1 - 10 млн. |  |  |  |  | 100 |  |  |

Продолжение таблицы 1.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Рынок сбыта** | | | | | **Кадры** | | |
|  | **3.1** | **3.2** | **3.3** | **3.7** | **3.11** | **4.2** | **4.4** | **4.5** |
| 1 | 1 | 2 | г. Ульяновск | 0,75 | Средняя стоимость человеко-часа у нас - 650р./ч. | 75 | 25 | 4 |
| 2 |  |  |  |  |  | 95 | 7 7% | 70 |
| 3 |  |  |  |  |  | 100 |  | 1 |
| 4 |  |  |  |  |  | 90 |  |  |
| 5 |  |  |  |  |  | 80 | 0 | 28 |
| 6 | 2898365 | 1 | Весь мир | 0,9 | Затраты на разработку программного комплекса составят 4 миллиона рублей. | 90 | 5 | 10 |
| 7 |  |  |  |  |  | 97 | 1 чел 3 % | 30 |
| 8 |  | 3 |  |  |  |  |  | 3 |
| 9 |  |  |  |  |  | 90 | 0 |  |
| 10 | 90 |  |  | 1/50 | 600 рублей в час | 90 | 0 | 11 |
| 11 |  |  |  |  |  |  |  |  |
| 12 | 100 | 150 | Россия, США | 0.94 | 1000 рублей / 1 час | 90 | 0 | 16 |
| 13 |  |  |  |  |  | 80 |  | 30 |
| 14 | 214 | 142 |  | 100 | 2000 руб/человекочас | 83 | 2,3 | 119 |
| 15 | 21 |  | Димитровград |  | 800 руб. час | 100 % |  | 1 |
| 16 | 124 | 35 | Москва, Петербург, Карелия, Новосибирск, Алтай, Магадан, Воронеж. |  | 700 рублей | 100 |  | 1 |
| 17 |  |  |  |  |  | 100 |  | 8 |
| 18 |  |  |  |  |  | 100 |  | 8 |

Продолжение таблицы 1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Внутренние процессы компании** | | **Инновационная деятельность** | **Соц программы** | |
| 1 | **5.7** | **5.8** | **6.1** | **7.6** | **7.7** |
| 2 |  |  | 30 | 0 | 0 |
| 3 | 1 | 5 |  | 0 | 0 |
| 4 |  |  | 100 |  |  |
| 5 | в США | в США |  |  |  |
| 6 | 7 | 2 | 0 | 0 | 0 |
| 7 | 5 | 2 | 100 | - | - |
| 8 | есть в США, 1 чел | есть в США , штатных сотрудников 2 + фрилансе количество меняется |  | 0 | 0 |
| 9 | 0 | 0 |  |  |  |
| 10 | 5 | 10 |  |  |  |
| 11 | 1 | 2 | 0 |  |  |
| 12 |  |  |  |  |  |
| 13 | 1 | 2 | 0 | 0 | 0 |
| 14 |  |  |  |  |  |
| 15 | 3 сотрудника | 5 сотрудников |  |  |  |
| 16 | нет |  |  |  |  |
| 17 | В отделе маркетинга работают 1 сотрудника. | В отделе продаж работают 1 сотрудника. | 10 |  |  |
| 18 |  |  |  |  |  |

Нами обрабатывались анкетные данные 87 ИТ-предприятий по 39 параметрам. Т.е. исходная матрица данных имела размерность 87×39. В таблице 1 приведены данные по нескольким предприятиям и по части параметров. Полный вариант таблицы 1 см. на лазерном диске.

**Входные данные программы.** Этой таблице поставим в соответствие матрицу, элементами которой являются данные из таблицы. Далее будем говорить о матрице и её элементах .

Данные в таблице 1 представлены как числовыми значениями, так и другими типами данных, как слова, текст и др., поэтому необходимо приведение данных к одному типу (однородность данных).

**Приведение данных к однородному виду.** Приведение данных к однородному виду производится отдельно в каждой группе.

1) Область деятельности компании.  
2) Финансово-экономические показатели.  
3) Рынок сбыта.  
4) Кадровые показатели.  
5) Внутренние процессы компании.  
6) Инновационная деятельность.  
7) Социальные программы и образовательная деятельность.

Специфика группы параметров «Область деятельности компании» заключается в том, что в группе один параметр, но он текстового типа данных.

**Таблица 2.**

Матрица преобразований и значений

параметра «Область деятельности компании».

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Назва-ние компа-нии | Обла-сть деяте-льно-сти компа-нии | Pазра-ботка | Продажа | Ремонт и обслу-жива-ние, сопро-вождение | Интер-нет и телекоммуникации | Обучение | Науч-ные иссле-дования | Защи-та информации | Значе-ние | Нормирован-ные значе-ния |
| ИнтелСофт | Заказ-ная разра-ботка про-грам-много обеспечения | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0,27128 | 0,444 |
| … | … | … | … | … | … | … | … | … | … | … |
| ФГУП "ЦентрИнформ" | Заказ-ная разра-ботка про-грам-много обеспечения | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0,02128 | 0,035 |
|  | Кол-во | 47 | 36 | 36 | 12 | 4 | 3 | 2 |  |  |

**Таблица 3.**

Общий вид матрицы преобразований и векторов , .

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Назва-ние компа-нии | Обла-сть деяте-льно-сти компа-нии | Pазра-ботка | Продажа | Ремонт и обслу-жива-ние, сопро-вожде-ние | Интер-нет и телекоммуникации | Обучение | Науч-ные иссле-дования | Защита информации | Зна-чение | Нормирован-ные значе-ния |
| ИнтелСофт | text0 | a0,0 | a0,1 | a0,2 | a0,3 | a0,4 | a0,5 | a0,6 | zn0 | nzn0 |
| … | … | … | … | … | … | … | … | … | … | … |
| ФГУП "ЦентрИнформ" | textn | an,0 | an,1 | an,2 | an,3 | an,4 | an,5 | an,6 | znn | nznn |
|  | Кол-Во | 47 | 36 | 36 | 12 | 4 | 3 | 2 |  |  |

Рассмотрим способы заполнения и назначение этих таблиц.  
Дан вектор состоящий из элементов (лингвистические переменные), , – номер предприятия по таблице 1. Необходимо его преобразовать в вектор , состоящий из элементов (числовые дробные значения), .

Преобразование осуществляется с помощью матрицы преобразования , состоящей из элементов (одно из двух числовых значений, либо либо ), j , – номер предприятия по таблице 1, – номера признаков (сфер деятельности), перечисленных по горизонтали в таблице 3.

Матрица преобразования строится по следующему принципу. Выделены несколько основных областей деятельности, которые теперь являются лингвистическими переменными: разработка; продажа; ремонт и обслуживание, сопровождение; интернет и телекоммуникации; обучение; научные исследования; защита информации. Проведён анализ текстовых записей на предмет принадлежности к области деятельности ( – если компания обладает признаком , – если нет).

Вектор значений строится с помощью матрицы преобразования по следующей формуле:

в знаменателе – количество ИТ-предприятий, обладающих признаком ; – весовой коэффициент для признака . Наибольший весовой коэффициент приобретает тот род деятельности, которым занимается меньше всего предприятий. – интегральный параметр, отражающий сферы деятельности -го предприятия, получаемый как сумма параметров строки таблицы 3 с соответствующими весовыми коэффициентами.

В этом случае каждый вектор-столбец таблицы делится на его - норму. Для полученных векторов-строк тоже ищется – норма. Это один из возможных способов ввода весовых коэффициентов. На практике часто для экономических моделей используют экспертные оценки весовых коэффициентов.

После получения вектора значений получим вектор нормированных значений по формуле:

.

Элементы получившегося вектора выражают, какую долю составляет параметр по отношению к максимальному параметру столбца, т.е. долю от показателя предприятия-победителя. Нормировка производится со следующими целями: чтобы все данные при кластеризации имели одинаковые единицы измерения (в долях от параметров-победителей), для уменьшения разброса данных, для приведения значений всех параметров в промежуток от 0 до 1. Здесь для нормирования векторов использована супремум-норма.

Во всех оставшихся группах (финансово-экономические показатели, рынок сбыта, кадровые показатели, внутренние процессы компании, инновационная деятельность, социальные программы и образовательная деятельность) параметры преимущественно числовые: тысячи рублей, проценты, да/нет(/). Их достаточно просто привести к нормированным числовым значениям.

Построим отдельно для каждой группы матрицу значений , состоящую из элементов (числовые дробные значения), ; , – номер предприятия, – номер признака внутри каждой из заявленных выше групп признаков. Перейдем к матрице нормированных значений , состоящей из элементов (числовые дробные значения), ; , вычисляемых по формуле:

.

Элементы получившейся матрицы выражают, какую долю составляет параметр по отношению к максимальному параметру -ого столбца, т.е. долю от аналогичного показателя предприятия-победителя. Нормировка производится со следующими целями: чтобы все данные при кластеризации имели одинаковую размерность (в долях от параметров-победителей), для уменьшения разброса данных, для приведения значений всех параметров в промежуток от 0 до 1. Здесь для нормирования векторов использована супремум-норма.

**Анализ данных.** После приведения данных к однородному виду мы получили матрицу данных . Ее элементами являются параметры дробного вида от до , включительно.

Если значение параметра равно , то будем считать, что параметр в анкете не был заполнен при анкетировании.

Построим матрицу наполненности анкетных данных по формуле:

.

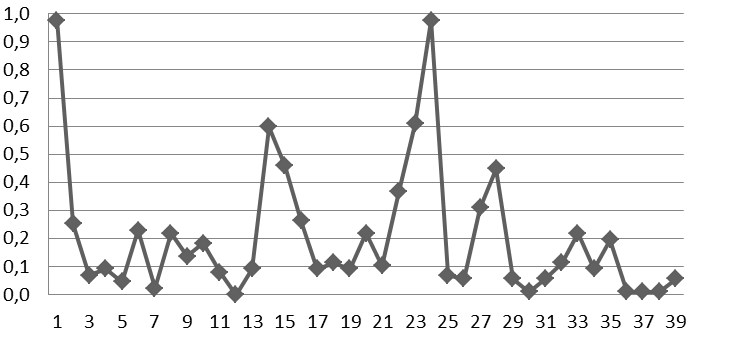
По матрице построим вектор «плотности заполнения анкет по параметрам» , где – номер вопроса анкеты по таблице 1,

где в числителе – количество заполненных предприятиями -ых вопросов анкеты,– плотность заполнения этого пункта анкеты 87 предприятиями, участвующими в анкетировании, в нашем случае

Построим вектор «плотность заполнения анкет по предприятиям» , где – номер предприятия по таблице 1.

в числителе – количество заполненных вопросов анкеты -ым предприятием, – плотность заполнения, в нашем случае .

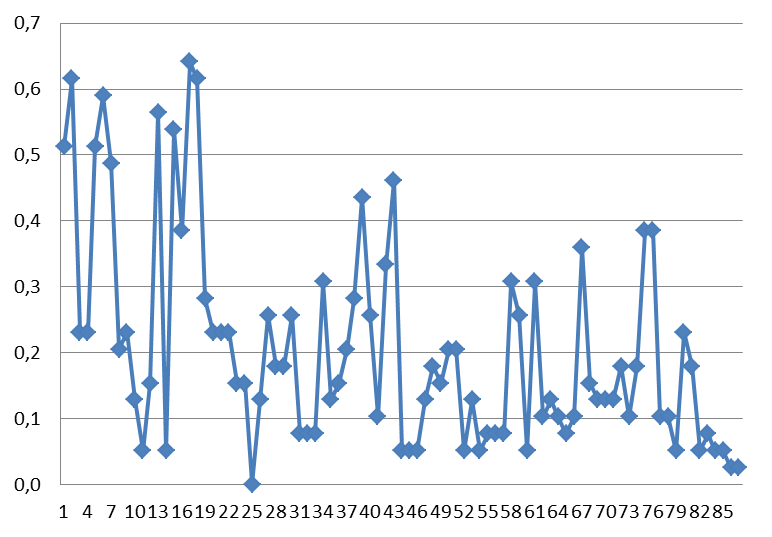
Построим график плотности заполнения 39 пунктов анкеты.



Пункты анкеты

Плотность заполнения пункта анкеты

Рис. 4. График плотности заполнения 39 пунктов анкеты.



Номер предприятия

Плотность заполнения анкеты предприятиями

Рис. . Плотность заполнения пунктов анкеты 87 предприятиями.

Рис. 5. Плотность заполнения анкеты каждым из 87 ИТ-предприятий

Воспользовавшись этими рисунками, можно получить представление о наполняемости анкеты ответами и выбрать уровень плотности заполнения строк и столбцов таблицы данных, выше которого данные можно считать заполненными. Данные ниже выбранного уровня можно обнулить, соответствующие строки – для предприятий, столбцы – для параметров исключим из таблицы значений.

**Редукция данных.** В нашем же случае «редукция» – сокращение объема данных. Редукцию можно провести на основе результатов анализа данных.

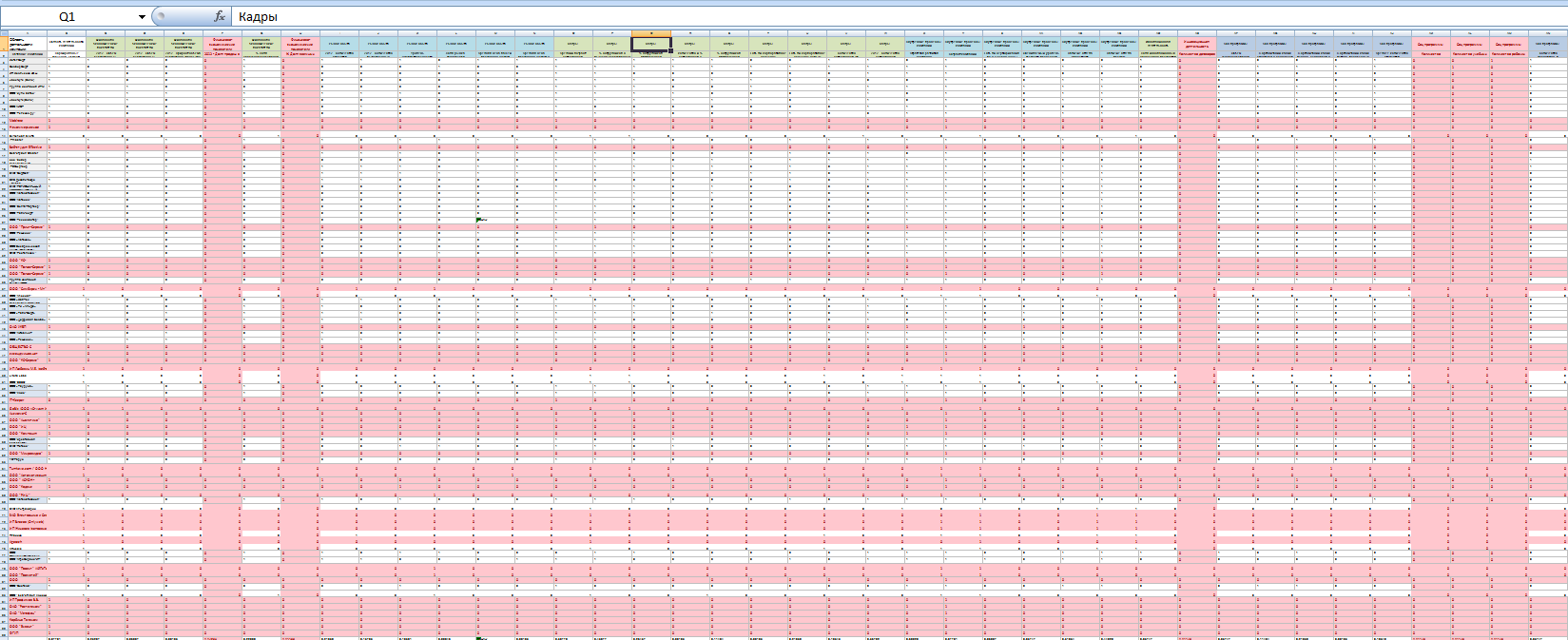


Рис. 6. Скриншот таблицы 1. Красным отмечены предприятия и данные с недостаточной плотностью заполнения (ниже выбранного уровня).

В результате редукции количество обрабатываемых предприятий сократилось с 87 до 49 , количество обрабатываемых параметров сократилось с 39 до 33. В результате мы получили матрицу данных размером и готовы приступать к кластеризации данных.