Маркетинг и маркетинговые исследования. – 2011. – №3(93). С. 184–192.

# КОМПЬЮТЕРНАЯ ТЕХНОЛОГИЯ ОБРАБОТКИ КЧЕСТВЕННЫХ ДАННЫХ ОПРОСОВ ПОТРЕБИТЕЛЕЙ ТУРИСТСКИХ УСЛУГ

***Н.С. Мартышенко**, к.э.н., доцент кафедры «Маркетинга и коммерции», Владивостокского государственного университета экономики и сервиса, профессор РА*

***Аннотация:*** В работе предложена методика обработки неструктурированных данных, которая существенно расширяет возможности исследователей при изучении социально-экономических явлений и процессов. Методика содержит в себе экспертную систему, использующую три вида компьютерных словарей. Представление информации в виде составных признаков позволяет производить анализ взаимодействия различных структурных групп потребителей

***Ключевые слова:*** методика обработки неструктурированных данных, экспертная система, анализ данных, туризм, анкетный опрос

Одной из ключевых проблем фундаментальной и прикладной науки в области туризма является разработка эффективных моделей прогнозирования и регулирования туристских потоков. Основным препятствием на пути их разработки и использования является отсутствие достаточной информации о процессах потребления туристского продукта на территории региона и тенденциях развития эволюционирующих потребностей человека в разнообразных видах рекреационной, досуговой, оздоровительной и других видах деятельности. Такая информация может быть получена в процессе массовых анкетных опросов потребителей туристского продукта.

Более восьми лет назад авторами была начата работа по проведению мониторинга поведения потребителей туристского продукта Приморского края. За этот период была проведена цела серия анкетных опросов.

При составлении текстов вопросов анкет выяснилось, что ряд важнейших характеристик процесса потребления туристского продукта не может быть определен при использовании обычных структурированных вопросов. Как показали многочисленные опыты опросов, человек более точно (и с меньшими затруднениями) отвечает на вопросы качественного или сравнительного характера, чем количественного.

Поэтому наши анкетные данные отличает то, что они содержат большое количество нечисловой информации, которая порождается использованием в анкетах разнообразных измерительных шкал. Наличие разнообразных шкал вызвано не прихотью исследователей, а их стремлением получить от респондентов более достоверную информацию. Исследователь всегда вынужден искать компромисс между желаемой информацией и информацией, которую он может получить в результате анкетного опроса. При составлении анкет нам пришлось широко использовать различные формы открытых вопросов.

Открытые или неструктурированные вопросы являются наиболее сложными с точки зрения компьютерной обработки анкетных данных. В отличие от закрытых, такие вопросы не содержат подсказок, не «навязывают» тот или иной вариант ответа и рассчитаны на получение неформализованного мнения. Многие исследователи не применяют компьютерную обработку открытых вопросов, а используют их в поисковых целях для получения информации для будущих исследований. Между тем, ответы на эти вопросы могут оказаться очень информативными.

Проблема компьютерного анализа качественных данных привлекает большое количество исследователей во всем мире. Определение проблемы компьютерного анализа качественных данных дано в работе [2]. Среди российских ученых, специализирующихся в области обработки качественной информации можно назвать Давыдова А.А. [3], Каныгина Г.В. [4].

Давыдов А.А. в качестве вывода к в своей статье написал, что «Современная российская социология катастрофически отстала от зарубежных достижений в области Qualitative Research, особенно в разработке

компьютерных систем Qualitative Data Analysis и Data Mining, не говоря уже об инновационных компьютерных информационных технологиях …» [3].

Каныгин Г.В. в работе [5] на вопрос «Стоит ли создавать свой «качественный автопром.?» отвечает «Конечно, стоит. Иначе просто невозможно осуществить тот самый инновационный путь развития.». Он утверждает, что «самым эффективным способом вхождения в искомую спецификацию является создание собственной компьютерной системы в своей предметной области».

Известные компьютерные средства обработки качественных данных, такие как NVIO [6] достаточно сложны для большинства маркетологов, а для российских исследователей еще и малодоступны. Поэтому нами была предпринята попытка создания простого компьютерного средства, которое могло бы быть доступно для освоения самому широкому кругу исследователей.

В настоящей работе мы предлагаем сначала рассмотреть элементы предложенной компьютерной технологии, позволяющей осуществить переход от неструктурированной формы представления информации к структурированной (типологии). Затем рассмотрим примеры построения типологий и использование их для анализа структур пространственного развития туристского комплекса региона.

**ОСНОВНЫЕ ЭЛЕМЕНТЫ КОМПЬЮТЕРНОЙ ТЕХНОЛОГИИ ОТКРЫТЫХ ВОПРОСОВ АНКЕТ**

Данные анкетного опроса принято представлять в виде таблицы «объект - свойство». Такую таблицу легко разместить на отдельном листе EXCEL. Для данных по открытому вопросу, представленных в форме текста, используется один столбец таблицы. Причем, мы считаем, что ответ может быть множественным. Например, отвечая на вопрос «Чем еще любите заниматься во время отдыха на море, кроме солнечных ванн и купания?» респондент может указать несколько вариантов ответа: « играть в волейбол, любоваться природой, ловить рыбу» и т.д. Признак в таблице «объект - свойство», содержащий данные по такому вопросу, мы называем составным.

Ответы в составном признаке могут состоять из нескольких более простых высказываний. Простые высказывания в ответах респондентов разделяются каким-либо знаком, например, «;» или «,». В более сложных случаях отдельные высказывания могут быть в виде целых предложений. В простейших случаях ответ может состоять только из одного простого высказывания. Допуская такие ответы на открытые вопросы, мы ни в чем не ограничиваем респондента.

Теперь определим, что мы имеем на выходе разработанной информационной технологии. Начнем с простых высказываний – частный случай составного признака (свойства). При открытой форме вопроса можно было бы ожидать, что респонденты не дадут одинаковых ответов. На практике встречается достаточно много одинаковых или сходных по смыслу высказываний, не говоря уж о простых описках и орфографических ошибках. Перечень действительно различных по сути, а не по форме, ответов на такие вопросы анкет ограничен. Уже при выборке порядка 700 анкет можно выделить всего от 30 до 50 по сути различных вариантов ответов, которые можно интерпретировать, как значения признака, измеренного в номинальной шкале. При увеличении объема выборки список вариантов практически не изменяется.

Для обработки данных открытых вопросов мы используем метод типизации. Метод типизации - это замена исходного простого высказывания (в форме текста) на близкое или сходное по значению, или обобщающее высказывание (в форме текста). Для выполнения операции типизации формируется вспомогательная таблица – «список значений признака». При расчете таблицы «список значений признака» составной признак разделяется на простые. Один из столбцов такой таблицы включает все уникальные значения исходного признака. Кроме того, она содержит столбец, в котором рассчитаны частоты встретившихся значений. Операция типизации применяется не к исходным данным таблицы «объект-свойство», а к данным таблицы «список значений

признака». Вначале обрабатываются простые ситуации. Например, различное написание одного слова или различный порядок слов. Среди сходных высказываний выбирается наиболее удачная (более точная или грамотная) форма написания высказывания, затем такое высказывание копируется в ячейки таблицы «список значений признака» со сходными высказываниями. Выполняя замену какого-то уникального высказывания на уже существующее из списка значений, мы, тем самым, сокращаем количество строк таблицы «список значений признака». После выполнения серии замен целесообразно выполнять операцию «сжатия», которая заключается в пересчете таблицы «список значений признака». Постепенно таблица «список значений признака» сокращается и становится более наглядной.

После того, как простые ситуации обработаны, приступают к обработке более сложных случаев. В таблице «список значений признака» отыскивается группа редко встречающихся, но касающихся одной темы высказываний. Для этой группы простых высказываний исследователь подбирает в таблице некоторое обобщающее высказывание, и если такого не находит, то сам формулирует новое обобщающее высказывание, отражающее общий смысл или тему группы простых высказываний.

Например, отвечая на вопрос «Чем еще любите заниматься во время отдыха на море, кроме солнечных ванн и купания?» наряду с другими ответами различные респонденты давали такие ответы:
«гонки на водном мотоцикле», «кататься на дельтаплане», «кататься на парашюте за катером», «прыжки в воду со скал», «скалолазание» и т.п.

Но эти высказывания встречались достаточно редко (менее 0,1%), поэтому мы заменили их на обобщающее высказывание - «экстрим», которое нашли в таблице «список значений признака». В принципе, смысл высказываний сохранился.

Чтобы не потерять информацию, особенно при повторном проведении опросов, мы заменяем сходные высказывания на обобщающие с уточнением. Уточнение или нюанс указывается в скобках. Например, в рассмотренном выше случае, мы заменили оригинальные значения на: «экстрим (гонки на водном мотоцикле)», «экстрим (кататься на дельтаплане)», « экстрим (кататься на парашюте за катером)», «экстрим (прыжки в воду со скал)», «экстрим (скалолазание)».

Для нас важнее характер ответа, который определяет тип личности респондента (потребителя), а не конкретное содержание ответа. Если исходная таблица «список значений признака» может содержать до нескольких тысяч значений, то после обработки (типизации) такая таблица обычно содержит до трехсот значений с учетом значений с уточнениями. Созданием такой таблицы заканчивается первый этап типизации (первый уровень).

Полученный новый признак содержит все еще слишком много различных значений, чтобы его можно было анализировать. Поэтому этот признак подвергается дополнительной обработке (второй уровень). На этом этапе просто исключаются уточнения, содержащиеся в скобках, и формируется еще один столбец таблицы «список значений признака», который мы называем подкласс, количество уникальных высказываний в котором будет уже от 30 до 50.

Наличие 30-50 вариантов значений - тоже большое количество для анализа измерений в номинальной шкале. Поэтому исследователь после формирования приемлемого списка действительно различных вариантов ответов, должен сгруппировать эти ответы, рассматривая их как некоторые характеристики непересекающихся классов, типов или тем, в зависимости от содержательного смысла признака и постановки задачи, для которой производится типизация. В нашем примере больше подходит определение типа личности. Объединение простых высказываний в классы является третьим уровнем типизации. Для каждого класса исследователь сам формулирует название по характеру объединяемых высказываний.

На практике результаты группировки у разных исследователей получаются очень похожими. Различия могут возникать из-за того, что некоторые высказывания действительно могут занимать промежуточное состояние и могут быть отнесены сразу к нескольким классам. Различия в группировке высказываний могут быть обусловлены различием критериев, которые используют различные исследователи для группировки высказываний. Названия классов каждый исследователь может дать совершенно разные. Лучше использовать лаконичные названия.

Таким образом, в результате обработки данных открытого вопроса мы будем иметь (на выходе):

- три новых представления признака (свойства), которые включаются в исходную таблицу данных и могут быть подвергнуты дальнейшей обработке для получения содержательных выводов;

- таблицу «список значений признака», которая может быть использована при повторении данного анкетного опроса или для выявления типизаций данных других анкет, которые предназначены для исследования данного процесса.

Необходимо отметить, что в результате типизации составных признаков будут сформированы также составные признаки. Для их анализа разработаны специальные методы обработки.

Технология обработки открытых вопросов имеет еще один важнейший результат, позволяющий существенно (на порядок) уменьшить время на типизацию данных при повторных опросах (мониторинге процесса). При пополнении таблицы исходных данных необходимо опять повторять процедуру типизации с учетом ввода новых данных. Для ускорения работы исследователь может использовать два типа словарей, которые создаются для каждого признака, содержащего данные по открытому вопросу: «Словарь замен» и «Словарь ключевых слов». Такие словари формируются для каждого отдельного качественного признака. Кроме того при обработке данных используется еще один словарь, который работает с различными качественными признаками и даже с различными анкетами. Это «Словарь избыточной информации». Он используется на первом этапе обработки качественной текстовой информации. С помощью этого словаря удаляются или корректируются высказывания, содержащие различную избыточную и несодержательную информацию.

Все словари хранятся в одном файле Access. Словари хранят опыт, накапливаемый исследователем в процессе работы по типизации высказываний и представляют собой базу знаний.

**РАЗРАБОТКА ТИПОЛОГИЙ ПОТРЕБИТЕЛЕЙ ТУРИСТСКОГО ПРОДУКТА**

Одним из наиболее простых примеров использования технологии является типизация наименований зон отдыха, посещаемых жителями региона в летнее время, например, Приморского края. В анкету «Исследование пляжно-купального отдыха» было включено два таких открытых вопроса:

- Чаще всего посещаю пляж: _____ (название бухты, острова или ближайшего населенного пункта)

- Чаще всего посещаю зону отдыха: _____ (название ближайшего населенного пункта, бухты или острова)

Первый вопрос относится к отдыху «без ночевок» второй - «с ночевками». При ответах на вопросы респонденты могут указать несколько излюбленных для посещения ими пляжей или зон отдыха. Список возможных мест, посещаемых во время отдыха, очень велик, а полный список вообще не может быть составлен самим исследователем. Поэтому для выявления структуры пространственного распределения зон отдыха может быть использован только открытый вопрос. При кажущейся видимости простоты задачи на практике она не так уж и проста. Практика показывает, что различных способов написания названий может быть очень много, при этом ошибки часто повторяются.

В качестве критерия группировки зон отдыха можно, например, использовать принадлежность к какому-либо административному району. После обработки плохо структурированных данных информация приобретает вид структурированной. В данном случае даже потерь информации не происходит. Исходная информация просто приводится в порядок, т.е. структурируется.

Рассмотрим более сложную ситуацию. Например, в анкете «Исследование пляжно-купального отдыха» нас интересовали предпочтения потребителей при организации своего времяпрепровождения в пляжной зоне отдыха. Ведь отдыхающие на морском побережье кроме принятия солнечных ванн и купания занимаются чем-то еще (особенно при отдыхе на побережье более одного дня). Для изучения времяпрепровождения отдыхающих в анкету был включен вопрос: «Чем еще любите заниматься во время отдыха на море, кроме солнечных ванн и купания: _____».

Как оказалось, спектр интересов отдыхающих не так уж и широк. После типизации высказываний они были объединены в 8 групп:

1. Спортсмены – 26%
2. Инертные – 19%
3. Увлеченные - 16%
4. Гурманы - 13%
5. Лирики - 12%
6. Общительные - 8%
7. Сони - 5%
8. Мамы – 1%

В той же анкете исследовались негативные высказывания респондентов по отдыху на море. Для анализа отрицательных мнений в анкету был включен вопрос: «Что омрачало ваш отдых в пляжной зоне: ». В результате типизации была определена вполне устойчивая структура распределения отрицательных реакций потребителей. После группировки отрицательных высказываний было выявлено 9 групп рекреантов:

1. Зеленые – 53%
2. Привередливые – 12%
3. Оптимисты – 10%
4. Нелюдимые – 8%
5. Урбанисты – 6%
6. Интеллегенты – 5%
7. Нетерпимые – 2%
8. Автомибилисты – 2%
9. Студенты – 1%

Методика компьютерной типизации также была использована при анализе ряда характеристик времяпрепровождения отпусков и каникул.

В анкетном опросе «Исследование туристского потенциала Приморского края и перспектив его развития» была использована наиболее сложная форма открытых вопросов. Это специализированная анкета, основу которой составляют вопросы, предполагающие ответы в форме нескольких предложений. Примером такого вопроса является вопрос: «Каким условиям должен отвечать городской пляж, при которых вы стали бы посещать пляж чаще, чем в настоящее время?». Даже при такой сложной форме вопроса с помощью метода типологии были выявлены вполне определенные группы различных мнений жителей Приморского края. Частотный ряд распределения мнений по сгруппированным данным представлен на рис. 1. Из диаграммы на рис. 1 следу-

ет, что большинство респондентов наряду с другими условиями, в качестве основного условия посещения пляжа выдвигают необходимость улучшения санитарного состояния пляжей (32%).

Это вполне обоснованные числовые оценки, над которыми необходимо задуматься и региональным органам управления, и коммерческим структурам, обслуживающим туристов и отдыхающих, чтобы организовать обслуживание потребителей с учетом изменяющихся потребностей туристов в разнообразных видах досуговой, оздоровительной, познавательной и других видах деятельности, являющихся основой и стимулом предпринимательского толчка к поиску и разработке новых региональных продуктов и видов туризма.

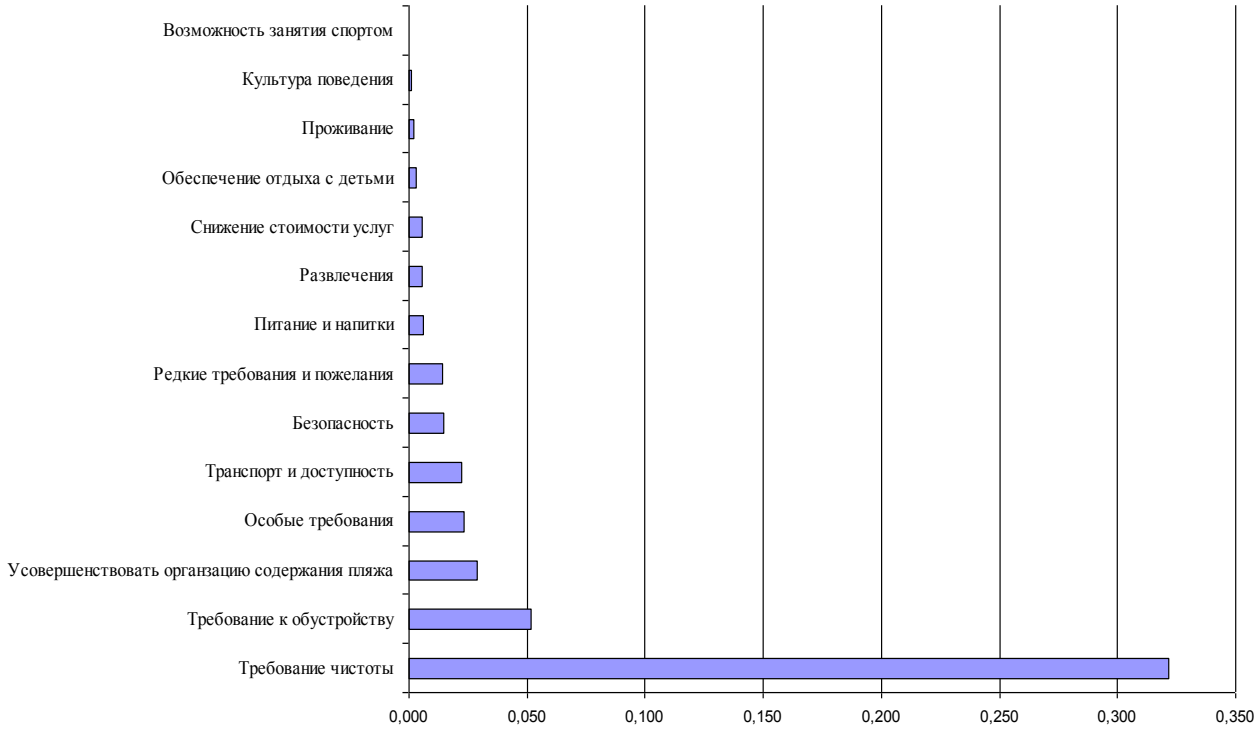


Рис. 1. Группировка высказываний потребителей по улучшению состояния пляжных зон отдыха

Мы рассмотрели три типа открытых вопросов. В первом случае ответ предполагается в форме одного или нескольких слов, во втором случае - это одна или несколько простых фраз, а в третьем случае - это сложные предложения.

**АНАЛИЗ СТРУКТУР ПРОСТРАНСТВЕННОГО РАЗВИТИЯ ТУРИСТСКОГО КОМПЛЕКСА РЕГИОНА НА ОСНОВЕ ТИПОЛОГИЙ**

Любые рекреационные ресурсы предполагают привязку их к конкретному месту или определение их пространственного расположения. Для прогнозирования и регулирования процессов потребления туристских ресурсов необходимо научиться описывать процессы с помощью структурированных данных. Ряд важнейших структурных характеристик мы получаем в процессе разработки типологий.

Характеристики структуры потребления вступают во взаимодействие и изменяются во времени. Целью стратегического управления является направить эти изменения в нужном направлении. А для этого необходимо разработать методики анализа структурных изменений.

Оказывается, что в чистом виде объекты далеко не всегда могут быть описаны одной структурной характеристикой. Например, по своим высказываниям, отдельные респонденты могут быть отнесены сразу к нескольким классам. Например, отвечая на вопрос «Что омрачало ваш отдых в пляжной зоне отдыха?», респондент может дать ответ: « экологическая обстановка, скопление людей, необустроенность пляжа» и т.д. Отдельные простые высказывания при типизации признака были отнесены к различным классам. В данном случае,

при типизации исходные значения были заменены на такие названия классов: «Зеленые», «Нелюдимые», «Урбанисты». Таким образом, по всей совокупности высказываний классы могут пересекаться. Однако оказывается, что одни классы более близки, другие – менее, а третьи – вообще изолированны. Рассмотрим принцип расчета оценки пересечения классов.

Для каждого составного ответа, заданного в форме классов, выделяются сочетания пар классов. Например, если отдельный ответ имел вид: «С, И, И, С, М» (где буквами обозначены классы), то респондент по своим высказываниям может быть отнесен сразу к трем классам. По данному ответу может быть составлено три пары сочетаний классов: «И, С», «С, М», «И, М». Таким образом, по возможным парам пересечений классов может быть составлена матрица пересечений. Размерность матрицы $к \times к$, где $к$ – количество выделенных классов. Матрица симметрична относительно диагонали. Элементы матрицы – это сумма встретившихся пар классов во всей выборке. Для исключения влияния на оценки пересечений классов объема выборки и размера классов элементы матрицы нормируются путем деления строк на количество пар высказываний по классам. Диагональные элементы характеризуют степень обособленности отдельных классов. Например, по сгруппированным данным относительно предпочтений времяпрепровождения в пляжной зоне была рассчитана матрица, представленная в табл. 1.

*Таблица 1*

Матрица пересечений классов

| Типизациия | Спортсмены | Увлеченные | Сони | Гурманы | Лирики | Инертные | Общительные | Мамы |
|---|---|---|---|---|---|---|---|---|
| Спортсмены | 0,708 | 0,089 | 0,013 | 0,087 | 0,088 | 0,003 | 0,069 | 0,003 |
| Увлеченные | 0,141 | 0,714 | 0,014 | 0,070 | 0,082 | 0,000 | 0,039 | 0,005 |
| Сони | 0,067 | 0,044 | 0,698 | 0,149 | 0,102 | 0,003 | 0,048 | 0,010 |
| Гурманы | 0,177 | 0,090 | 0,060 | 0,608 | 0,069 | 0,005 | 0,110 | 0,005 |
| Лирики | 0,178 | 0,105 | 0,041 | 0,069 | 0,631 | 0,003 | 0,060 | 0,006 |
| Инертные | 0,004 | 0,000 | 0,001 | 0,003 | 0,002 | 0,993 | 0,001 | 0,000 |
| Общительные | 0,210 | 0,075 | 0,029 | 0,165 | 0,090 | 0,002 | 0,563 | 0,012 |
| Мамы | 0,079 | 0,079 | 0,048 | 0,063 | 0,079 | 0,000 | 0,095 | 0,667 |

Для наглядности пересечений классов их связи удобно изобразить в виде графа (рис. 2).

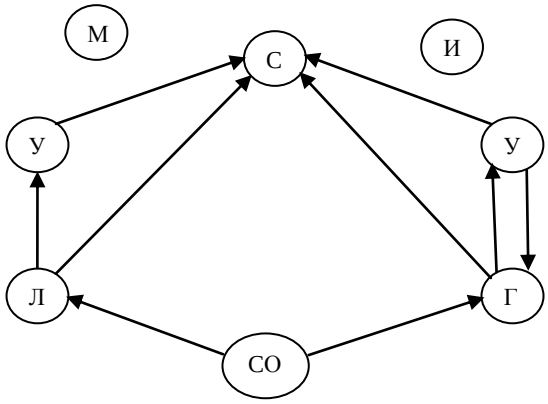


Рис. 2. – Граф пересечений классов

При построении графа устанавливается некоторое пороговое значение на оценки пересечений. Тогда на графе будут присутствовать только наиболее существенные связи. Граф на рис. 2. построен с пороговым значением 1,5. Классы в вершинах графа обозначены по первым буквам их названий. Граф позволяет судить о том,

как необходимо сочетать спектр предлагаемых услуг, т.е. разрабатывать дизайн регионального туристского продукта для различных групп потребителей.

**ЗАКЛЮЧЕНИЕ**

1. В работе предложена методика обработки неструктурированных данных, которая существенно расширяет возможности исследователей при изучении социально-экономических явлений и процессов. Основными преимуществами компьютерной технологии являются простота, доступность и эффективность. Исследователь, занимающийся обработкой анкетных данных, в состоянии освоить основы технологии в течении одного сеанса. Технология обработки качественных данных разработана как надстройка к EXCEL, что обеспечивает ей полную доступность. Эффективность работы подтверждена длительным сроком эксплуатации в течении которого было решено множество задач.

2. Методика содержит в себе экспертную систему, основанную на использовании трех видов компьютерных словарей. С помощью этих словарей различные исследователи, занимающиеся в одной области, могут производить обмен накопленным опытом по систематизации информации.

3. Представление информации в виде составных признаков позволяет производить анализ взаимодействия различных структурных групп потребителей.

4. Предложенная методика обработки данных прошла апробацию при обработке значительных объемов реальных данных [7,8] и может быть рекомендована для использования широкому кругу исследователей, применяющих в своей практике анкетные опросы.

5. Предложенная методика является шагом в направлении анализа качественных данных, собираемых через Интернет (Социология 2.0). Сегодня проблема больше даже не в обработке данных, а в сборе данных и внесении их в компьютер, а при сборе данных через Интернет этот процесс можно автоматизировать.

**Литература**

1. Мартышенко Н.С. Методика сбора и обработки данных для оценки структуры потребителей услуг туристского комплекса региона // Практический маркетинг. — 2009. — №11. С. 16 – 28.

2. Lewins A., Taylor C., Gibbs G.R. What is Qualitative Data Analysis, 2005, http://onlineqda.hud.ac.uk/Intro_QDA/what_is_qda.php (считано 03.02.2010).

3. Давыдов А. А. Качественные исследования: перспективы развития. М.: ИС РАН, 2008. (http://www.isras.ru/index.php?page_id=922)

4. Каныгин Г.В. Конструируя конструктивизм//Социол. исслед. 2006, № 11, С. 19-28.

**5.** Каныгин Г.В. Ответ А.Давыдову http://www.isras.ru/index.php?page_id=926

6. Explore. Discover. Share http://www.qsrinternational.com/#tab_you

7. Мартышенко Н.С, Власенко А.А. Анализ стратегий развития туристского комплекса региона // Территория науки. — 2007. — №2(3). С. 175–182.

8. Мартышенко Н.С, Старков А.С. Анализ структуры потребительского регионального туристского рынка // Территория науки. — 2007. — №4(5). С. 468–478.