ДИАЛОГОВАЯ СИСТЕМА КЛАССИФИКАЦИИ И АНАЛИЗА ТЕКСТОВ[1]

Е.А.Каневский, Г.И.Саганенко, Л.М.Гайдукова, Е.Н.Клименко

(Москва)

ДИСКАНТ представляет собой систему для классификации и обработки как текстовой, так и другого рода анкетной информации, которая хранится в базе данных системы. Обсуждаются новые особенности системы для анализа открытых вопросов. ДИСКАНТ открывает широкие возможности для анализа текстовых данных в социальных науках.

Ключевые слова: системы анализа данных, базы данных, контент-анализ, открытые вопросы, классификация текстов, итеративная классификация, анализ анкетной информации, статистическая обработка данных, визуализация данных.

Анализом содержания текстов занимаются многие исследователи при изучении влияния средств массовой информации на общественное мнение, при изучении документов истории и культуры, при изучении политического, экономического, юридического и даже экологического сознания общества. Одной из первых систем автоматизированного анализа текстов является **General Inquirer** (Гарвард, 1968), основанная на широком использовании различных словарей [1]. Современная система **TACT** (Торонто, 1990) позволяет вычислить отношение встречаемости данного слова в окрестности выбранной точки (ключевого понятия) к общей встречаемости этого слова. Наиболее развитой из известных является система **TEXTPACK** (ZUMA, Center for Survey Research and Methodology, Mannheim), которая также основана на широком использовании словарей [2].

Немного теории

Одним из методов качественно-количественного изучения содержания текстов является *контент-анализ* (КА). В процессе КА все многообразие текстов по интересующей исследователя тематике сводится к набору определенных элементов, которые затем подвергаются подсчету и анализу. В отличие от лингвистического анализа при КА подсчитывают не лингвистические единицы, а элементы содержания, которые можно определять по-разному, чем и вызвана некоторая субъективность результатов.

Обычно в качестве элемента содержания (единицы анализа) при "машинном" КА используют слово, которому ставят в соответствие определенную категорию. Это удобно, так как слово выделено в тексте пробелами изначально. Иногда для обозначения категории используют два-три слова, образующих устойчивое понятие. Однако слово характеризуется лишь номинативной, назывной функцией. Единицей выражения мысли является предложение, которое используется в качестве единицы содержания при классическом ("ручном") КА. Мы применяем в качестве элемента содержания несколько другую единицу анализа - *фразу*, которая может состоять как из целого предложения, так и из нескольких слов, и даже из одного слова. Каждая фраза является выражением одного суждения, одной мысли. При

[1] Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований. Грант №96-0680216.

автоматическом сравнении фразы считаются идентичными друг другу при совпадении двух-трех слов или одного-двух ключевых слов, более точную оценку идентичности фраз может дать исследователь при полу автоматическом анализе [3].

Анализируемые тексты имеют определенную структуру и представляют собой множество суждений, формулировок по реализации, поименованию некоторой идеи, состояния, явления и т.п. В частности, в социологии такой материал дают суждения экспертов, высказывания по темам интервью, ответы на открытые вопросы и др. В психологии - это формулировки законченных предложений в тестах. В средствах массовой коммуникации - это названия публикаций и передач, их тематическая направленность.

Любые попытки применения КА к текстовым массивам неизбежно связаны с проблемой классификации и, следовательно, с разработкой *классификатора*. В большинстве исследований заранее составляется формализованная, полностью закрытая схема классификации еще до начала обработки материалов [4]. В результате этого получается стандартное описание текстовых массивов. При достаточно часто повторяющихся исследованиях, например, в средствах массовой информации, стандартный или, по крайней мере, почти постоянный классификатор даже помогает сравнивать результаты последнего обследования с предыдущими.

При анализе быстро меняющейся ситуации даже опытный исследователь, проводя обследования через 1-2 года, не может заранее, до получения материалов опроса, создать полностью готовый классификатор. Причина ясна: сместилась тематика ответов респондентов, их волнует уже не то, что год или два тому назад. Поэтому на один и тот же вопрос (открытого типа) они отвечают совсем не так, как раньше. В этом случае классификатор должен уточняться непосредственно в процессе КА, при осмыслении материалов данного опроса. Естественно, что необходимо обеспечить достаточно простой способ коррекции и пополнения классификатора, а также возможность сравнительно простой переориентировки фраз из одной группы в другую, из одного класса - в другой. Иначе все попытки изменения классификатора не дадут результатов.

Следует отметить, что простейшая структура классификатора обычно напоминает таблицу и содержит категории (группы) и их модальности (типы). При этом каждая категория имеет один и тот же набор модальностей. Более сложный классификатор имеет древовидную структуру, состоящую из классов и групп, причем каждый класс может иметь свой набор групп.

Система ДИСКАНТ

Основополагающей для диалоговой системы классификации и анализа текстов (**ДИСКАНТ**) является идея упростить работу с текстовой информацией, используя методы КА текста совместно со словарным анализом. Система является оригинальным программным продуктом и не имеет прямых аналогов.

Система **ДИСКАНТ** предназначена прежде всего для обработки произвольных текстовых ответов на вопросы структурированной анкеты в сочетании с "жесткой" (количественной) информацией закрытых вопросов. Кроме того, система позволяет обрабатывать неструктурированную текстовую информацию разного рода, а также выполняет некоторые элементы анализа текста: составление словарей и указателей, подсчет частоты встречаемости слов, поиск слов в тексте и в словаре и т.п.

В состав системы входит встроенная реляционная *база данных* (БД), которая обеспечивает хранение цифровой и текстовой информации в формате, близком к формату DBF-файла. Созданы специальные средства для облегчения разработки классификатора фраз и работы с ним. В рамках общего массива информации система позволяет вести обработку целого ряда текстовых подмассивов, соответствующих различным темам и требующих автономных процедур классификации и анализа. Результаты можно вывести на экран,

распечатать или записать в файл. Полная визуализация результатов облегчает пользование системой.

Основные характеристики версии 5.4

Выбор режимов работы системы осуществляется с помощью оконных меню.

•***Окно СТРУКТУРА*** обеспечивает задание структуры БД и ее редактирование. БД может содержать до 300 полей, в каждом из которых целесообразно хранить ответ респондента на один вопрос. Имя поля, соответствующее вопросу анкеты, содержит до 44 символов (как русского, так и латинского алфавита). Имеются поля символьного, целого и смешанного типов (анкеты с открытыми, закрытыми и полузакрытыми вопросами). Они имеют размер до 234 символов, при необходимости поле автоматически расширяется до 1248 символов. Поля других типов (числовые, логические, даты и др.) не могут расширяться.

•***Окно БАЗА ДАННЫХ*** обеспечивает возможность заполнения БД путем ручного ввода данных, их просмотра и коррекции по полям и по анкетам. Каждая анкета заносится в БД в виде самостоятельной записи, размер которой не превосходит 32 Кбайт. Хотя по своим техническим параметрам БД допускает до 2 млн. записей, для последующей словарной обработки необходимо, чтобы произведение количества анкет на количество полей не превышало 40960. Обеспечивается возможность выделения ключевых слов, разбиения текстовых ответов респондента на фразы и отбор некоторых из них для первичной классификации. Предусмотрена возможность формирования новой БД на основе уже имеющихся.

•***Окно СТАНДАРТНЫЙ СЛОВАРЬ*** обеспечивает возможность создания словарей слов, ключей и фраз по текстам, хранящимся в БД. Размер словаря - до 150 Кбайт, каждый элемент словаря имеет ссылку на номер анкеты и поля. Словари позволяют анализировать частоту встречаемости отдельных элементов и осуществлять различного вида поиски. В частности, имеется возможность поиска элементов в стандартном словаре или фраз в тексте (по словам из словаря) при задании на поиск всего слова или его начала. Можно осуществить также глобальный поиск, при котором последовательно перебираются все слова из словаря и для каждого выводится соответствующий текст из БД.

•***Окно КЛАССИФИКАТОР*** обеспечивает возможность создания классификатора, его просмотра и редактирования. Сам классификатор представляет собой древовидную структуру имен классов и групп, он может иметь до 25 классов, по 99 групп каждый. Обеспечена возможность классификации фраз с образованием базовых фраз для последующего анализа. Принадлежность любой фразы к той или иной группе (и классу) может быть изменена в процессе вторичной классификации.

•***Окно НОРМАТИВНЫЙ СЛОВАРЬ*** обеспечивает возможность создания нормативных словарей слов, ключей и фраз из базовых фраз. Каждый элемент такого словаря имеет ссылку на номер соответствующей фразы. Нормативные словари используются для КА в процессе идентификации фраз.

•***Окно АНАЛИЗ*** обеспечивает возможность идентификации фраз, то есть сравнения всех фраз из выбранного поля с базовыми. Возможен анализ распределения фраз из заданных полей по классам и группам. При наличии управляющей БД обеспечивается анализ по вторичным признакам с выводом результатов в виде двумерных гистограмм. Кроме того, предусмотрена возможность анализа распределения количества фраз и сочетаний классов.

•***Окно ОБРАБОТКА*** обеспечивает возможность статистического анализа ответов респондентов на закрытые вопросы анкеты. Вычисляются одномерные, двухмерные, трехмерные и многоальтернативные статистические анализы.

•***Окно СЕРВИС*** обеспечивает возможность импорта данных из текстового файла в БД и обратно (в нескольких форматах). Для уточнения размеров полей при заведении новой БД можно получить распределение по длине текстовых ответов.

База данных системы

Система имеет гибкую в обращении базу данных (БД), позволяет хранить первичную информацию в удобном и естественном виде, обеспечивая простой доступ к массиву в целом, анкетам, полям (вопросам), их редактирование, просмотр, вывод на печать или в файл. Рассмотрим некоторые особенности БД.

•В обычных случаях работы с БД пользователь, как правило, легко мирится с сокращенными наименованиями полей. Если же в БД занесены анкеты и система в целом должна обеспечить их обработку, то все время возникает необходимость соотносить ответ респондента (содержимое поля БД) с вопросом анкеты (наименованием этого поля). Увеличенный по сравнению с обычным размер наименования поля и возможность использовать в нем как русский так и латинский алфавит обеспечивают достаточно адекватные наименования полей.

•Кроме стандартных для БД полей [5], в рассматриваемой БД имеются поля целого и смешанного типов. Такой набор типов полей позволяет наиболее адекватно хранить ответы респондентов, учитывая, что в анкетах имеются открытые, закрытые и полузакрытые вопросы, причем последние могут быть как одно-, так и многоальтернативные.

В поле типа "Целое" могут размещаться одно или несколько целых чисел, которые соответствуют номерам вариантов ответов, выбранных респондентом. Они разделяются друг от друга запятыми.

В поле типа "Смешанное" могут размещаться одно или несколько целых чисел и текстовый ответ респондента. Различаются два подтипа таких полей. В первом случае за числами следует текст, во втором - за текстом следуют числа. В обоих случаях числовая часть специальным знаком отделяется от текстовой. Это позволяет производить обработку разных частей такого поля независимо друг от друга и в то же время полностью сохранить структуру анкеты.

•База данных системы обладает еще одной характерной особенностью: поля могут иметь переменную длину. Это дает возможность пользователю просмотреть несколько анкет и определить по ним примерную длину полей. Если при вводе ответов респондента они не помещаются в данное поле, то достаточно включить расширение и в распоряжении пользователя оказывается почти весь экран - 1248 символов.

После окончания набора данного ответа респондента формируется новое поле с учетом фактической длины ответа, а весь текст размещается в двух местах: начало в заданном пользователем поле, а окончание - в специальном файле расширения, причем в поле об этом делается соответствующая отметка. Если поле имеет расширение, то просмотреть все содержимое такого поля можно только в режиме редактирования.

•С помощью встроенных в систему вспомогательных программ созданная любым образом (путем импорта текстового файла или ручным набором текста с помощью системы) база данных легко может быть реорганизована, причем могут быть изменены как размеры полей, так и порядок их расположения.

Если пользователь хочет изменить размеры полей с целью сократить размер БД или обеспечить максимально возможный вывод информации из поля, то он может экспортировать массив анкет в текстовый файл. Затем с помощью вспомогательной программы можно определить максимальные размеры всех полей, а при желании и увидеть распределение длин ответов по каждому полю в отдельности. Эта же программа определяет для каждого поля его оптимальную длину, то есть такую длину, при которой около 75% ответов не требуют расширения. Если задать рекомендуемые размеры полей и импортировать текстовый файл в новую БД, то как показывает анализ анкет, полученные при этом размеры файлов БД близки к минимально возможным.

Возможности системы

•После создания структуры БД и ввода информации в систему все исходные тексты разделяются на фразы, каждая из которых сфокусирована на одной теме. Прописными буквами выделяются ключевые слова. Массив анкет в дальнейшем обрабатывается или последовательно анкета за анкетой (продольный разрез), или по одноименному полю всех анкет (поперечный разрез). Кроме того, путем наложения ряда условий может быть задан подмассив (подмножество) анкет для последующей обработки. С помощью аппарата стандартных словарей выполняется предварительный анализ выбранного подмассива текста, в результате чего наиболее характерные и повторяющиеся фразы отбираются для *первичной классификации*.

•На основе анализа текстового подмассива и предыдущего опыта исследователь создает первичный вариант классификатора. Он может создаваться для каждого поля в отдельности, а может объединять несколько полей.

•Далее производится классификация ранее отобранных фраз, в процессе которой каждой фразе присваивается соответствующие ее смыслу класс и группа. Если это затруднительно, то временно можно не задавать группу (и класс) - такой фразе присваивается нулевая группа (и нулевой класс). Проделав эту процедуру со всеми файлами отобранных фраз, которые должны быть объединены в данном классификаторе, мы получим массив базовых фраз. В процессе классификации происходит уточнение и дополнение самого классификатора.

Таким образом, классификация выполняется внутри системы и не требует априорных схем. При этом для осуществления более оптимальной классификации используется несколько приемов поддержки: стандартные словари, режимы глобального поиска, поиск фраз по словам и пр.

•На основе базовых фраз формируются *нормативные словари* слов, ключевых слов и фраз, после чего по каждому полю отдельно выполняется процедура *идентификации текста.* Она осуществляется путем контент-аналитического сравнения всех фраз с базовыми в диалоговом режиме. Если для очередной фразы найден аналог, то его идентификатор фиксируется в файле идентификаторов. Если нет, то фраза отправляется в файл дополнительных фраз. После окончания процедуры идентификации (а она может быть прервана в любой момент) такие фразы классифицируются дополнительно и пополняют массив базовых фраз, затем идентификация повторяется. Такой итерационный процесс продолжается до полной идентификации всего текстового подмассива.

Сущность предлагаемой методики сводится к тому, что вместо 100-процентной классификации всех фраз проводится классификация только части фраз, а все остальные фразы отождествляются с ними. Это дает двойное преимущество. Во-первых, сокращается объем работы по классификации фраз. Во-вторых, при любом изменении классификатора - а это не исключение, а правило при анализе текстов - достаточно изменить класс и группу у ряда базовых фраз, чтобы автоматически произошли соответствующие изменения и у всех остальных фраз, аналогичных им.

•Наиболее просто анализируется *распределение фраз по классификатору*. Результаты выводятся в процентах в виде таблицы, гистограммы или круговой диаграммы. Можно сопоставить результаты итоговой классификации для одного и того же классификатора по любым возможным подмассивам.

Для сопоставления текстовых и числовых ответов следует сформировать вторичные признаки в *управляющей БД*, которая имеет ту же структуру, что и основная БД. Для каждого признака запоминается его имя, номера используемых им полей, условия его вычисления и ряд служебных параметров. После этого система позволяет анализировать распределение фраз по классификатору и вторичным признакам с представлением результатов в виде таблицы и "двумерных" гистограмм. Для таких гистограмм характерно неравномерное

распределение ответов по классам и признакам. Одна из подобных гистограмм, на которой представлено распределение 1955 фраз по двум признакам (позитивные и негативные ответы) и 10 классам, изображена на рисунке.

Для заданного набора полей можно получить распределение фраз по их количеству. Это позволяет выяснить, какие вопросы больше волнуют респондентов, вызывают больше эмоций.

Результаты анализа сочетаний классов в каждой анкете для выбранного набора полей представляются в виде двух многомерных таблиц. В первой из них приводится число всех встречающихся сочетаний классов. Все возможные комбинации из них приведены во второй таблице. Этот анализ позволяет понять, сочетания каких тем встречаются в ответах чаще всего.

•Таким образом, полный цикл анализа каждого текстового подмассива состоит из пяти этапов: создание массива отобранных фраз, разработка классификатора, классификация отобранных фраз и формирование массива базовых фраз, идентификация оставшихся (не включенных в базовые) фраз с базовыми, получение распределений. Каждая итерация позволяет пополнять базовые фразы и классификаторы и соответственно изменяет результаты идентификации и распределения. На любом этапе можно проверить результаты классификации и идентификации и вернуться назад, исправив данные или принятые исследователем решения.

•*Анализ числовых данных* (жестких признаков) является традиционным для систем обработки анкет. Статистическая обработка данных в рассматриваемой системе производится только для полей целого или смешанного типа, соответствующих закрытым или полузакрытым вопросам. Прежде всего необходимо определить максимальное количество рангов (градаций) для каждого из этих полей и закрыть полузакрытые вопросы. Затем надо ввести нужное количество рангов для каждого из полей, после чего можно приступать к статистическому анализу ответов на одноальтернативные вопросы:

- при одномерном статистическом анализе результаты выводятся в виде таблицы и гистограммы, кроме того, подсчитываются значения средних величин и энтропии;

- при двумерном статистическом анализе результаты выводятся в виде таблицы, подсчитываются значения χ^2, средних градаций обоих полей, коэффициентов сопряженности и корреляции Пирсона;

- при трехмерном статистическом анализе результаты выводятся в виде многомерной таблицы, в которой для каждого имеющегося набора рангов выводятся количество анкет и их процентное содержание.

При многоальтернативном статистическом анализе результаты выводятся в виде двух многомерных таблиц. В первой из них для заданного набора полей приводится количество всех встречающихся сочетаний рангов. Все возможные комбинации из них представлены во второй таблице.

Некоторые особенности ДИСКАНТ

•Если информация попадает в БД системы путем импорта из текстового файла, то легко обеспечить правильность набора путем предварительной проверки текста на какой-либо системе контроля правописания. Обычные текстовые редакторы имеют строки длиной до 255 символов, системы контроля правописания работают со строками той же длины: **ДиаКор -** со строками до 250 символов, **Корректор -** до 252 символов. Это достаточно хорошо согласуется между собой.

Но как быть, если информация вводилась в **ДИСКАНТ** вручную, и отдельные ответы респондентов превышают 255 символов? Для решения этой проблемы имеется специальный режим экспорта данных с автоматическим разбиением длинных строк на несколько корот-

ких, причем обеспечен последующий импорт такого файла в БД системы после его проверки и коррекции.

•Система позволяет анализировать и отдельные тексты, вообще не имеющие какой-либо структуры. Для анализа такого текста, набранного в одном из текстовых редакторов в формате ASCII, следует создать БД, состоящую из одного поля символьного типа и перед импортом файла переформатировать исходный текст на максимально возможную длину строки, исключив переносы. При желании разместить в одной записи несколько строк их следует заключить в "абзацные" скобки (знаки "<" и ">").

•Хотя ДИСКАНТ формально и обеспечивает работу с текстами, набранными латинскими буквами, вся система настроена на обработку русскоязычных текстов. Для эффективной работы с текстами на других языках требуется определенная языковая настройка.

•В заключение отметим, что для работы системы **ДИСКАНТ** достаточно IBM AT 80286, MS-DOS 5.0 и 0,8 Мб свободного места на жестком диске или дискете.

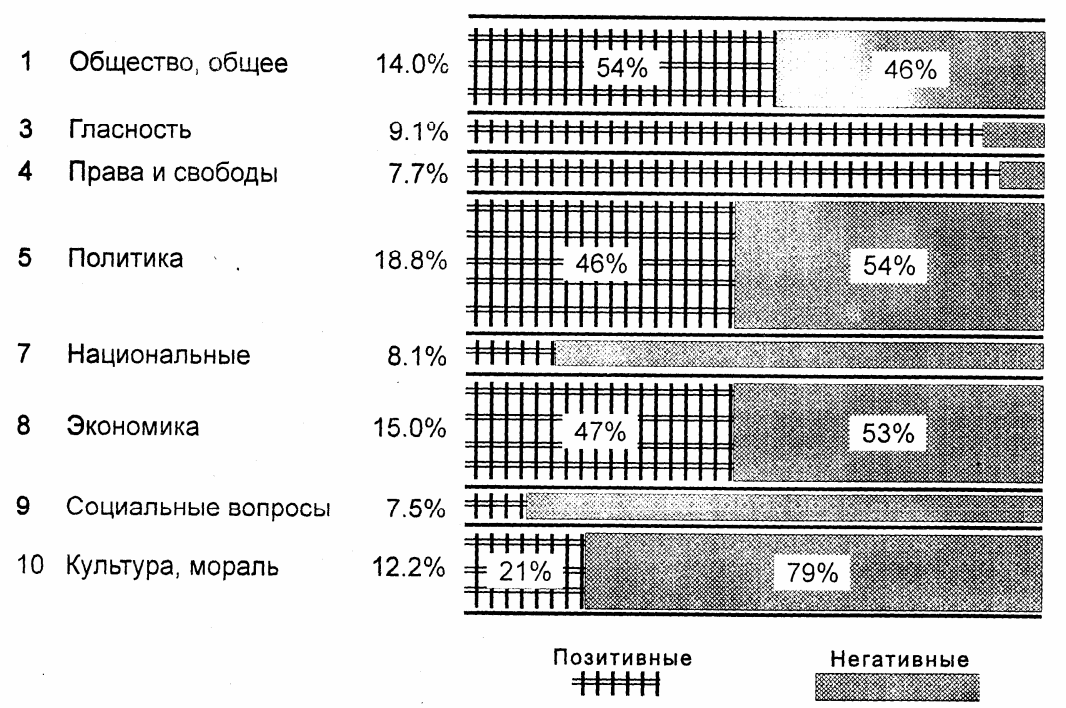

Рис. 1. Распределение 1955 фраз по 10 классам и 2 признакам

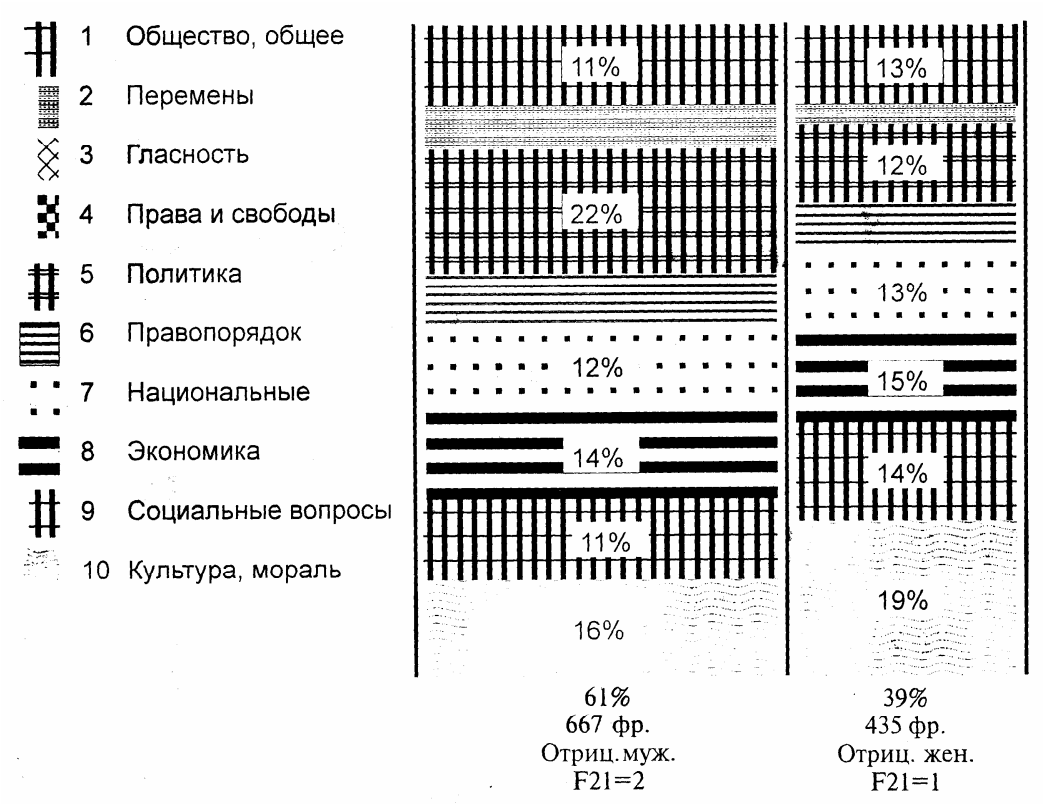

Рис. 2. Распределение фраз по двум группам и 10 классам

ЛИТЕРАТУРА

1. *Coxon A.R.M., Trappes-Lomax H.R.N.* INQUIERER III (Edinburg's version). Edinburg Univer., Jan.1977, Rep.92.
2. Text Analysis and Computers//Conference Programme and Abstracts. Mannheim, Germany: ZUMA, Center for Survey and Methodology, 1995.
3. *Каневский Е.А., Клименко Е.Н., Гайдукова Л.М.* Контент-анализ текстов и проблемы идентификации//Информационные технологии в гуманитарных и общественных науках. СПб:Спб ЭМИ РАН, 1995.
4. *Коробейников В.С.* Методы качественно-количественного анализа содержания документов//Методы анализа документов в социологических исследованиях. М.:ИСИ АН СССР, 1985.
5. *Гринберг Ф., Гринберг Р.* Самоучитель программирования на входном языке СУБД dBASE III/Пер. с англ. М.:Мир, 1989.